# A CATEGORICAL DATA ANALYSIS OF HEALTH PRACTICES, HEALTH STATUS, AND HOSPITAL UTILIZATION IN METROPOLITAN ST. JOHN'S

CENTRE FOR NEWFOUNDLAND STUDIES

TOTAL OF 10 PAGES ONLY
MAY BE XEROXED

(Without Author's Permission)

BARBARA M. VEITCH

# A Categorical Data Analysis of Health Practices, Health Status, and Hospital Utilization in Metropolitan St. John's

By

©Barbara M. Veitch

A practicum report submitted to the
School of Graduate Studies in partial fulfillment
of the requirements for the degree of
Master of Applied Statistics

Department of Mathematics and Statistics
Faculty of Science
Memorial University of Newfoundland
February, 1991

St. John's Newfoundland Canada

**Abstract**

Questionnaires were administered to adults from a sample of households in the Metropolitan St. John's area to gather data on their lifestyles, health habits and utilization of medical care services.

Health practices, as described in the social medical literature (eating breakfast, smoking, drinking, sleeping, correctness of weight, and exercising), are explored. A variety of statistical measures of association are used to gauge the strength of the relationships between these variables and one's health status.

The relationships between sleeping habits and one's health is examined using logistic regression. This analytical technique is again employed to study the effect of alcoholic consumption on health and to further explore its effect once educational level is controlled for.

From individual health practices, a weighted health practice index is developed. Using loglinear analysis we build models so as to examine the association between this score and hospital utilization, controlling for sex, age and education.

**Acknowledgements**

Funding for the research was provided by the National Health Research and Development Program, Department of Health and Welfare, Canada.

The author gratefully acknowledges the assistance of Dr. Roy Bartlett, Associate Professor, Department of Mathematics and Statistics, who in his role as supervisor, provided guidance and advice with infinite patience throughout this study. Thanks are also extended to Dr. Jorge Segovia, the principal medical researcher on this research team and to the Division of Community Medicine for providing the opportunity to work on this project. Finally, I wish to express my gratitude to my family for their support and encouragement.

## Preface

The author joined this study after the proposal for the survey was accepted. She was employed as a research assistant for the duration of the study and was responsible for assisting in the pre-testing and revision of the questionnaire and the hiring and training of interviewers. As a member of the field office and research team she was involved with editing questionnaires, quality control and data cleaning in addition to the other tasks required of this team. As well, she was responsible for the initial analysis of data.

During the course of the author's involvement with the study, two presentation papers were written by the research team and presented by one of the principal investigators, Dr. Jorge Segovia – one to the American Public Health Association in June of 1986 and the other to the Canadian Public Health Association in September of the same year. The Federal Government Report was submitted in January, 1987.

# Contents

# List of Figures

# Chapter 1

# Survey Design and Sampling

## 1.1 Introduction

This was a study using a telephone survey, of lifestyles, health practices, and medical care utilization. It was designed in part to consider health indicators and how these indicators are related to health status and medical care utilization. A unique feature of this study was the linkage of the survey data with data pertaining to hospital discharges and physicians' services.

While there is some explanation of the execution of the survey from the standpoint of the field office work, in this report we will concentrate on some of the statistical issues - from the sampling procedure to the examination of design effects and the analysis of data. Data are studied primarily with association measures, logistic regression and loglinear analysis.

## 1.2 Sampling

One of the first things to address once it was decided what was wanted from the study, was the sample design.

### 1.2.1 Population

The population to which the survey results apply consists of all people 20 years of age or older in St. John's, Newfoundland. The sample was selected and the questionnaires were administered in the spring/summer of 1985. As will be outlined, there were restrictions placed on the population definition due to the sampling frame and the accessibility of some of the would-be respondents. Given that the limitations were not very severe, we need not be exceedingly cautious in generalizing to the population initially defined.

### 1.2.2 Frame

Essentially the frame for this study was one section of the Newfoundland and Labrador Telephone Directory published in March of 1985, immediately prior to the selection of the sampling units. The section of interest in the directory covered the St. John's area. It should be noted that the frame, as such, exceeds the frame of interest. For example, this section contained a small number of telephone numbers outside of the Metropolitan St. John's area. Because of this, the definition of Metropolitan St. John's was limited to those residences having St. John's exchanges as listed at the front of the directory. This in itself presented some difficulties since in certain areas some residences had these exchanges while others did not. It was questionable whether or not such areas should belong to the frame but it was felt that the definition was easiest to apply if held consistent for all St. John's exchanges. Any deviations were considered minimal and seldom occurred.

Using boundaries, the area of interest could have been defined geographically. This was ruled to be too time-consuming as many street addresses would have to be manually checked for the region to which they belonged. In defining the area by

exchange codes then, such neighboring places as Bay Bulls, Petty Harbour, and Outer Cove, for example, were sometimes included – 'sometimes' as in these places some homes had a St. John's exchange while others d'.l not. This was not regarded to be a serious problem as these households were not considered to be very different from those of St. John's, per se. Also, residents of these areas, being geographically very near St. John's, have the same medical facilities available to them. Such households appeared in the sample relatively infrequently.

As well as the above, the frame as defined exceeded the frame of interest in terms of private households in that it included telephone numbers such as those belonging to businesses and institutions.

There were two groups of people who were perhaps u.der-represented or excluded altogether. Elderly people are likely to be under-represented since old age homes were excluded. This should be qualified. Old age homes which house the elderly in self-contained apartments with private telephone numbers and which provide minimal nursing care were included. The elderly are probably further underrepresented in that many of the non-respondents in the study *were* non-respondents because they were hard-of-hearing and since the survey was a telephone survey it would be especially difficult to interview such a person. It is perhaps a fair assumption that the majority of such persons were elderly. Those who were not well enough to answer the questionnaire were also excluded. Although our non-response rate was reasonably low, these people should be kept in mind together with those who, at the time of the survey, resided in an 'excluded institution' such as a hospital.

A group of people who were ignored altogether were those with no telephones or those with unlisted telephone numbers. According to the Newfoundland Telephone Company, telephone coverage in St. John's is approximately 99% of which about 4%

are unlisted telephone numbers. Given that the number of unlisted telephones is small, their exclusion from the survey is unlikely to bias the overall results. However, these persons would have had the same chance of being selected as those with listed numbers had random digit dialing been used. Since no automatic random digit dialing equipment was available, the procedure would have had to have been carried out manually. The number of 'useless' numbers generated by a computer program could be substantially reduced providing that only St. John's extensions were permitted for the first three digits, but non-existent numbers would be generated nonetheless. Also, many business numbers would be generated and would only be discarded once the number was called and the place was identified as such. Sampling methods for random digit dialing which reduce the number of useless telephone numbers in the sample have been proposed by Waksberg and Mitofsky (1978), for example. In a later paper, Potthoff (1987) generalizes their technique. Although considered, random digit dialing was not implemented as at this time no bank of numbers from which telephone numbers could be generated was available for release from the telephone company. We were comfortable with the assurance from the telephone company of almost total listed telephone coverage in the survey area.

Other excluded numbers, and hence possible would-be respondents, were those associated with prisons, hotels, and institutions such as hospitals.

Ideally our frame would have consisted of an enumerated list of all St. John's exchange telephones – preferably including unlisted numbers and excluding those not of interest such as government departments and businesses. Unfortunately the telephone company could not release such a list.

### 1.2.3 Sampling Procedure

Stratification of the population on a suitable variable may have been valuable. In some cases telephone exchanges are identical with some characteristics of the population and such exchange numbers may be used to stratify the population. But in the present survey, stratification by telephone exchange numbers is not related to any study variable of interest. Our procedure was to take a random sample of households as determined by telephone number selection.

The medical researcher was interested in collecting information on all adult members of a given household. A simple random sample was taken on households in the St. John's area and once a household was selected, all persons in that household aged 20 years of age or older were approached to be interviewed. That is, we took a single-stage cluster sample with clusters of unequal size (see Cochran 1977).

One sampling method which was considered was systematic sampling. Since names were to be selected from the physical telephone directory, this would have facilitated the task of actually selecting members for the survey. That is, only one random number would be generated from which point every $k^{\text{th}}$ household would be selected. In this way systematic sampling tends to distribute the sample over the population frame more evenly. In the same way, this can also result in periodicity. The listed population, however, was the telephone directory and it is likely that alphabetization alone does not group persons in any fashion; they are still randomly listed from the point of view of the study.

The nature of the variables in this study is such that most of the data are categorical. The intention was that contingency tables would be examined by means of measures of association and logistic and loglinear analyses. Given this, it was desirable to have as close to a simple random sample as possible. Hence, although

a systematic scheme of sampling would have made the job of selection somewhat easier, the need for a simple random sample from the analysis viewpoint outweighed this factor. The sampling procedure, however, was not a simple random sample of survey members but rather a simple random sample of clusters of unequal sizes, the average cluster size per sampling unit or household being approximately two people aged 20 years of age or older. What this meant in terms of violation of underlying assumptions of the analytical techniques shall be discussed in a later section on design effects.

### 1.2.4 Sample Size

To make a reasonable determination of sample size, it is advantageous to have some idea of which statistical methods will be employed in the analysis of data. The intention of this study was to model and test for association between variables in two-way and multi-way tables (Segovia et al. 1987). A commonly used test in this type of analysis is the $\chi^2$ test. Based on this type of test one can determine the sample size required provided one fixes the desired significance level, $\alpha$ (the probability of mistakenly rejecting the null hypothesis, $H_o$), the power of the test, $1-\beta$ (where $\beta$ is the probability of mistakenly accepting $H_o$), and the 'effect' size (an index of degree of departure from $H_o$).

If our hypothesis of independence between variables is correct, we would expect certain frequencies of occurrence, or proportions thereof, in each cell of the contingency table. Specifically, the proportion of occurrence in each cell would be $\frac{1}{m}$, where $m$ is the number of cells. If the experimental or observed contingency table exhibits the same proportions as that which would be expected under the null hypothesis, then we would not reject the hypothesis of independence or no association. The strength of association between variables is reflected in the degree of departure from

the expected proportion. Cohen (1977) uses $W$ to index the size of such departures, or effect size.

$$W = \sqrt{\frac{\lambda}{n}}$$

where $n$ is the sample size and $\lambda$ is the noncentrality parameter of the noncentral $\chi^2$ distribution. Cohen provides tables of sample sizes required for the analysis of contingency tables when $\alpha$, $1 - \beta$, and $W$ are fixed.

It was known at the outset of the study that several contingency tables would be analyzed and these were considered when choosing an appropriate sample size. One such table cross-classified the frequency of doctors' visits (broken down into three categories) with three levels of health practices. To determine the sample size, $\alpha$ was set at .05, the power at .80, and the effect size $W$, at .30 – a 'medium' value suggested by Cohen for contingency table analysis using $\chi^2$ tests. With these fixed, for a 3×3 table with 4 degrees of freedom, Cohen's tables give the sample size of $n$ =133.

In the Alameda County Study (Belloc and Breslow 1972, Belloc 1973, Breslow and Enstrom 1980) respondents were asked to report on a total of six health habits. The outcome was that 12.4% practised 0-3 of the health habits, 52.3% practised 4-5, and 35.3% practised 6-7. The Medical Care Plan (MCP) files from St. John's were examined to ascertain the marginal distribution of doctors' consultations. Of the patients who had a visit to a doctor, 63.7% had 1-5 visits, 20.4% had 6-10, and 15.9% had $\geq$11. It seems reasonable to use these marginal distributions as approximations for the distributions of these variables in our study. So then, this *a priori* information was used to calculate the expected values in the cells of the contingency table. These calculated proportions suggested that a 'large' effect may have sufficed. With $\alpha$=.05 and $1-\beta$=.80 as before, together with Cohen's suggested value of $W$=.50 for a 'large' effect, the sample size was $n = 48$.

It was decided that this contingency table be controlled for by sex and age. Let us think of our two-way contingency table, *doctors' consultations×health practices*, as being one layer of the four-way table *sex×age×doctors' consultations×health practices*. The 1981 census gave the marginal distribution for gender in St. John's as males 45% and females 55%. For age the marginal distribution was 56.2%, 27.6%, and 16.1% for 20-44 years, 45-64 years, and ≥65 years, respectively. All else being equal, with the addition of sex and age, the cell with the smallest expected proportion of occurrence for *doctors' consultations×health practices* would be that for males ≥65 years of age. It seems reasonable therefore, to make the following calculation to obtain a minimum required sample size.

$$\begin{aligned} n &= 133/(.45\times.161) = 1836 \quad \text{if} \quad W = .30 \\ \text{or}\quad n &= 48/(.45\times.161) = 663 \quad \text{if} \quad W = .50 \end{aligned}$$

Given that we anticipated a response rate of 90% and a 90% linkage rate with medical utilization files, the sample size was inflated to

$$\begin{aligned} n &= 1836/(.90\times.90) = 2267 \quad \text{if} \quad W = .30 \\ \text{or}\quad n &= 663/(.90\times.90) = 819 \quad \text{if} \quad W = .50 \end{aligned}$$

The following was also considered when determining the sample size. Theoretically, in the analysis of contingency tables using $\chi^2$ tests, the sample size should be sufficiently large so as to avoid cell frequencies that are too small. There is no definitive value, however, for 'small'. Fisher (1970) is one of many who recommend a minimum expected cell frequency of 5. The literature on the subject tends to use this value as an acceptable rule of thumb (Hays 1981, Freeman 1987, Kraemer and Thiemann 1987) although Hays, for one, suggests a minimum value of 10 in 2×2 tables. It is also put forward – particularly if the number of degrees of freedom is large – that provided no more than 1 out of 5 cells has a frequency of less than 5, a minimum expected frequency of 1 is permissable in these cells (see Hays, for ex-

ample). Camilli and Hopkins' (1978) empirical studies found that even with small expected cell frequencies in 2×2 tables, Pearson's $X^2$ test is very robust.

Considering the four-way table, *sex*×*age*×*doctors' consultations*×*health practices*, recall the marginal distribution of each of these categorical variables. The minimum proportion for each was as follows: sex, 45% (males); age, 16.1% (≥65 years); doctors' consultations, 15.9% (≥11 visits); health habits, 12.4% (0-3 habits). Hence, for a minimum expected value of 1, the sample size required is

$$\begin{aligned} n &= 1/(.45\times.161\times.159\times.124) \\ &= 700. \end{aligned}$$

To ensure a minimum expected frequency of 5 would require a sample size of

$$n = 5\times700 = 3500.$$

Recalling our anticipated response rate of 90% and linkage rate of 90%, we need a sample size of

$$\begin{aligned} n &= 3500/(.90\times.90) \\ &= 4321. \end{aligned}$$

With all the above calculations in mind and considering restrictions on time, cost and manpower, a sample size of 3000 was chosen.

### 1.2.5 Selection of Sampling Units

As we were to use a simple random sample to select our sampling units, a FORTRAN program was used to generate a series of random numbers. The numbers generated were associated with a given line of the directory in the section of interest. It was quite accurately estimated that approximately 50% of the numbers would correspond to non-residential numbers; therefore, the quantity of numbers generated was twice that which would be required for the sample. This program did the following:

- generated a random sample of given size from 72,333 (the number of lines in the St. John's section of the directory).
- printed these numbers together with their corresponding page, column, and column position in the directory.
- randomly assigned an equal quantity of these numbers to a given number of interviewers.

We originally intended on a sample size of 3000 individuals. Using the Statistics Canada figure of 2.3 adults 20 years of age and older per household in St. John's at that time, this translated into 1304 households. To obtain this we had to select 1304/.5=2608 lines. Prior to the commencement of the survey execution, this number was increased when reconsideration of our assumed response rate of 90% led us to decide that a rate of 80% would perhaps be more realistic. Therefore the number of lines to be selected by the program was recalculated to be 2608(.9)/.8=2934.

# Chapter 2

# The Survey Execution

## 2.1 Pre-testing

At this point the questionnaire was ready for pre-testing. Several people were selected at random from the telephone directory and the questionnaire was administered over the telephone to them. This was done to ensure that all questions were phrased in a way that was clearly understood and not ambiguous. As well, it was important to check that the layout of the questionnaire was logical and easy for the interviewers to follow. It was also necessary to check the length of time required to administer the questionnaire. Additional questions pertaining to salary and MCP number were included after the pre-testing.

The pre-testing suggested some minor alterations to the phrasing of some of the questions, and a couple of sections had the actual layout of the questions altered to make it easier for the interviewers to follow the question sequence. An additional section was placed at the end of the questionnaire. This section contained the information from the household sheet and the information regarding the total number of refusals and non-respondents in a given household which was to be filled out after the interview was completed.

## 2.2 Types of Households

Although the pre-testing was only done on individuals, the survey was to be carried out on all adult members of the selected households. Consequently a household sheet was required. Upon first contact with a household, the interviewer would be required to get a list of those people in the unit eligible to be interviewed and each person's relationship with the 'head' of the household. It was not possible in this survey to have a household sheet that could clearly categorize each type of dwelling which could be encountered. A form was constructed which would categorize households as accurately as possible without being so complicated that it would cause confusion or inconsistency on the part of the interviewers.

In the end we allowed for three types of households - a family household, a household of unrelated people, and a single adult household. Even then, of course, not every household could be expected to conform exactly to one of these set types. In a 'family' household, for instance, people living within the dwelling but unrelated to household members were not considered as part of the unit. In a case where two unrelated families were residing together, the family whose name was listed in the telephone directory was taken to be the selected 'family' household. Where a married couple was living with parents/in-laws and the name listed in the telephone directory was that of the younger married couple, then that couple constituted the 'husband' and 'wife' and the parents of this couple were entered as such. In a household of five residents, if only two were siblings then that household would be recorded as an 'unrelated' household and the siblings would not be recorded as related. On the other hand, if four of these people were siblings, they would become members of a 'family' household and the fifth person would not be considered as a member of that dwelling.

When it was unclear as to which category a household belonged, interviewers were instructed to contact the supervision office where a decision would be made and recorded so that in the event that similar households came into the survey, they would be classified consistently.

## 2.3 Training: The Interviewer's Manual

Once the questionnaire was finalized the next step was to write an Interviewer's Manual. It was comprised of information on the following:

1. Interviewing Skills :

   This briefly stressed the importance of the role of the interviewer in a survey.

2. Ethics of Interviewing :

   The duty of the interviewer to be discrete and ensure confidentiality was emphasized. Information obtained from respondents was to be disclosed to no one with the exception of supervisors.

   The importance of initiating and maintaining a comfortable but professional interaction was discussed. Interviewers were not to express approval or disapproval of a subject's response, nor were they to give leading probes such as "You do. .don't you?"

3. "Do's" and "Don'ts" of Interviewing :

   Primarily, this summarized in point form that which had already been mentioned. It also mentioned that interviewers were not to interview friends, acquaintance or relatives and highlighted some of the things to be kept in mind when editing completed questionnaires.

4. Field Work Procedures :

This included information on the number of households which would be allotted weekly to each interviewer and the number of individual interviews this would be expected to yield. It also informed the interviewers that weekly meetings would be held to assess progress, sort out problems, deliver completed questionnaires and collect new assignments. As part of standard practice, spot checks would be made by supervisors with the respondents of completed questionnaires.

Once households were assigned, the procedures were outlined for making contact and returning completed questionnaires. In order to make an initial contact with a household, interviewers were to make up to seven attempts on different days and times of the day – five calls within the first month and two in the next month. In the case of a refusal of an entire household or an individual within a household, a letter requesting participation was to be sent from the field office and the interviewer was to call back four working days later. At the end of each week, questionnaires from completed households, together with household sheets and interviewer record forms, would be turned into the field office. Interviewers were instructed to call the field office whenever they had a query or problem so that such queries would be handled immediately and consistently.

5. Completing the Household Sheet :

One household sheet was to be completed for each household. The three types of household classifications – family, unrelated, and single adult – were defined. Not every household would fall neatly into one of these categories and instructions were given as to what to do in this eventuality. In addition, an explanation was given regarding how to assign an identification number to each household

member.

6. Question Instructions :

This section addressed each of the 69 questions in the questionnaire. It clarified questions, explaining for example, that 'animal fats' include food such as dairy cream, table or 'real' butter, whole milk, fatty meat and gravy. It gave instructions on how to record answers and how to use probes.

7. The First Contact with the household :

This gave the initial statement to be used upon first contact with a household.

8. The Informed Consent Statement :

Given here was the informed consent statement which was to be read to each individual before commencing the interview.

9. Editing :

Included among the instructions regarding editing, interviewers were directed to:

- edit their questionnaires as soon as practicable following the completion of the interview, ensuring that every appropriate question was answered.
- transfer all information to the coding blocks on the questionnaires, using '9', '99' and so on if the question was inapplicable or if the subject did not know how to answer or refused to answer a question.

10. Questions the Interviewer Might be Asked :

A list of several questions that a respondent might ask, together with suggested responses, was included. If, for example, a subject expressed concern about the confidentiality of the study, the interviewer could respond by saying that

everything she is told is confidential, is seen only by the staff, and that no person is ever identified in any reports.

This guide was considered to be an important document to use in the training of interviewers and for their reference throughout the course of the field work.

Seven female interviewers, two of whom had previous interviewing experience in survey-type studies, were initially hired. Shortly after selecting these interviewers, a one-week training period was scheduled immediately prior to the commencement of the field work. During this week, the Interviewer's Manual was covered methodically to ensure that everyone understood the skills, ethics and so forth, involved in survey interviewing. Each item in the questionnaire was discussed.

Interviewers then practised administering the questionnaire on each other and edited and corrected each other's work. Queries were encouraged and discussed. Each interviewer was given a list of households which she was expected to contact over the course of two or three evenings. These were for practice only and not included in the analysis. These households had also been selected at random from the telephone directory but were not taken from the list of households to be used in the household survey. That is, they were selected independently of the survey sample (although checked to ensure there was no overlap). With these 'practice' households, questionnaires and household sheets were to be completed and editing was to be done immediately upon finishing each interview. Each day interviews completed during the previous evening were discussed among the group.

## 2.4 The Commencement of Interviewing

Once the training period was over, the survey started in earnest. Interviewers were instructed to complete as close to 40 questionnaires per week as possible. They were not to go beyond this quota since there were only two field supervisors who, among their other duties, were responsible for editing questionnaires after delivery to the field office. In addition, there was an upper limit on the number of questionnaires which would be entered onto the computer system each week at Newfoundland and Labrador Computer Services. It would be best if the interviewing were carried out at the same rate as the editing and data entry so that when errors occurred or clarification was required on a given questionnaire, this fact would be uncovered as close to the time of the interview as possible. This reason was twofold. First, if the interviewer herself could answer the query, she would be much more likely to be able to do so shortly after the interview than after a period of a week or more. Second, if a follow-up call to the respondent were required, it should be done as soon as possible.

### 2.4.1 Interviewers - Keeping Tabs

Originally it was intended that a certain percentage of the interviews would take place at the field office under direct supervision, but unfortunately it did not turn out to be viable. Physical space limitation was such that the only room available to us in which on-campus telephone interviews could be conveniently made, was only large enough to accommodate one interviewer with one supervisor. Although it was a disadvantage that interviews could not take place under direct supervision, it was hoped that other supervisory methods would suffice.

In addition to keeping track of the team's work through meetings and careful editing, some 'running tabulations' were kept. Each week and for each interviewer

the number of households was recorded, together with the number of people in each household less than 20 years of age, the number at least 20 years of age, and the number of refusals, non-respondents and respondents. All this information was obtained from the household sheets. From individual questionnaires several variables were recorded. With these few variables, some comparisons could be made between interviewers and with census information. We will discuss later a problem uncovered by these running tabulations.

Interviewers were compared for number of refusals, non-respondents, respondents and number of completed questionnaires. Small discrepancies in the number completed per week were both expected and accepted. Concern over differences in the number of questionnaires completed was not as great as that over differences in ratios of refusals and/or non-respondents with the total number of possible respondents from the households. For the most part, such ratios did not exhibit statistical differences between interviewers although some interviewers generally appeared to elicit more responses than others.

As well as comparing interviewers regarding the above, the research team was interested in the response rate itself since, of course, the projected response rate influenced the sample size. Also, regardless of the number of responses, it was obviously a concern that the refusal rate be as low as possible so as to reduce possible bias in the results.

The interviewers' distributions on variables such as sex, marital status, height, and number of people per household at least 20 years of age, were compared. Any consistent and significant differences between interviewers would warrant closer inspection. If a given interviewer deviated consistently from her co-workers, it might suggest that the questionnaire was not being administered in the way it was intended

or that short-cuts were being taken. Although the questions recorded were perhaps not the best to uncover if an interviewer were taking short-cuts, they still served their intended purpose to some degree. In large part, the reason for the choice of these questions among all the possible questions was simply that census data, while generally not readily available on most variables, were available on these. This also allowed the research team to check that the data from the sample selected was in keeping with census data on these variables specifically and, therefore, hopefully on other variables in general. In particular, the number of people per dwelling who were $\geq$ 20 years old was of interest; the census figure of 2.3 adults per household was used in calculating the number of households to select. A deviation from this could greatly influence the sample size since it was households and not individuals which were selected from the directory. Our average was slightly less than this and to compensate for the reduction in the number of possible respondents that this caused, we generated several more random numbers. It was assumed that the slight discrepancy only indicated a minor change in the population since the census of 1985 or a slightly different definition of a household for our survey than that used by the census. Hence, increasing the number of households to sample would not bias our results. The variable *sex* was of interest since the ratio of males to females was another factor in our choice of sample size. Knowing the sex was also important in that other statistics (such as marital status and height) were available in the census broken down by gender. As well, the variable height was not useful unless the sex of the respondent was known.

## 2.5 Data Entry, Processing, Checking and Cleaning

Newfoundland and Labrador Computing Services (NLCS) was approached in the early days of the study when the proposal for the project itself was being drafted.

Their services were employed for data entry on the understanding that they would receive approximately 200 questionnaires per week. Although we requested 40 questionnaires per week from each of our seven interviewers, we were correct in our assumption that we would not exceed this number on a weekly basis.

The coding area of the questionnaire was designed in consultation with their staff so as to maximize facility of data entry and hence reduce the number of data entry errors. In addition, they were to enter the information twice and flag any non-matching entries. A program also checked for a limited number of 'out-of-bounds' data points.

Each week when questionnaires were brought to NLCS, the previous week's work was collected and returned to the field office, together with any tapes onto which the data had been transferred. The tapes were then copied onto the university's computer system. Once there, programs were run to test whether the measurements on the aforementioned 'running tabulation' variables were statistically the same among the interviewers. These tests brought to light the rather disturbing fact that data from one of the interviewers were consistently and statistically differing from the others. This prompted the field office staff to make callbacks to a sample of respondents for each interviewer. Respondents were informed that this was a standard random check to ensure that the interviews had been conducted properly by the interviewers and they were requested to answer again a selection of the questions It quickly became evident that in the case of six of the seven interviewers the questions were being answered by respondents to the field office staff as they had been to the interviewers. For one of the interviewers, however, this was not so. Of course, one might expect and accept slight discrepancies between the first and second interview, especially if more than a week had passed, but such discrepancies were much more pronounced in the case of the one interviewer. Unfortunately this interviewer had to be dismissed.

The approximately 500 questionnaires which she had completed were redistributed among the other interviewers and readministered. A statement was prepared for the interviewers to read to these respondents explaining that it had been discovered that the questionnaire had perhaps not been carried out correctly in the first instance and requesting that they repeat it. These were completed again with surprising results; rather than refusing to repeat the interview or being aggravated by the request, the majority of these respondents were very obliging. In fact, many seemed pleased that the research team was being careful regarding their data; others were relieved, stating that they had not been impressed at the way the questionnaire had been administered in the first instance. The response rate was very good. In retrospect, the fact that the problem only became evident after several weeks makes it more clear that every effort should be made in the future to have at least a percentage of the interviews administered under direct supervision.

Although the response rate from these questionnaires was very good, it was important to check that they were not different from the other completed questionnaires. Several variables were tested for statistical difference between the repeated questionnaires and the others. When no significant differences surfaced, the research team was satisfied to pool the data from these questionnaires with the data from the others.

As the data became available to the research team, the data cleaning continued. Errors to be checked included those uncovered through the program which flagged errors during data entry, 'coding' errors such as a 3 being coded where there could only be a 1 or a 2, and 'logical' errors such as a person who reported having never smoked later stating that he smoked a package of cigarettes each day. Suspected outliers were also checked. The questionnaires from which the errors surfaced were examined. If the values on the questionnaire and in the data file corresponded but

were impossible or extremely unlikely, a callback was sometimes in order; otherwise, in the event of an impossible answer, the value was recoded as 'missing'.

## 2.6 Refusals and Non-respondents

It was anticipated that some household members would request additional information about the study before agreeing to participate or would require more information pertaining to the request for their MCP number. Hence two letters were drawn up – one explaining the nature of the study and the other justifying the request of MCP numbers. Both letters repeated the promise of confidentiality. For would-be respondents who refused to answer the questionnaire, two additional letters were prepared -- one for complete household refusals and one for individual refusals.

When any of these situations arose, interviewers were instructed to contact the field office immediately. From there the appropriate letter would be mailed. After several days the interviewer was to contact that household again. If the person still declined to participate in the study, the household sheet, together with any completed questionnaires from that household, was to be returned to the office. Once all the households in the survey were contacted and interviews completed, the refusals were pooled and redistributed among the interviewers. No interviewer was to receive her own refusals to readminister. This yielded good results with many people granting interviews to a different interviewer. Once this stage was complete, there was a 90% response rate among those households in which at least one person answered the questionnaire. The remaining 10% were not all refusals, per se, but rather some were 'non-respondents'. These included people who were perhaps too ill to come to the telephone, but this subgroup seemed to be largely made up of the hard-of-hearing.

With respect to not being able to make even an initial contact with a household

or an individual, interviewers were instructed to try at least seven times before regarding the household or individual as unobtainable. Sometimes one member of the household was temporarily absent so had to be contacted several days or even weeks after the initial contact. If he were to be gone for longer than this, he was considered unobtainable.

**Summary: Households for which there was ≥1 response**

| | Frequency | % of Total Number of Subjects |
|---|---|---|
| Households | 1675 | |
| Subjects | 3649 | |
| Respondents | 3300 | 90.4 |
| Refusals | 195 | 5.3 |
| Non-Respondents | 154 | 4.2 |

The above summary refers only to those households for which there was at least one response. These correspond to households for which the household sheet (which recorded the number of responses, non-responses and refusals) was completed. It ignores entirely the households where no response could be obtained. The sample listing consisted of 2076 households. Of these, 1675 had at least one respondent. Of the remaining 401 households, 179 were contacted and of these, 148 were complete household refusals and 31 were household non-respondents. The remaining 222 consisted of households for which the telephone number was no longer in service (N/S) or for which no contact could be made after seven attempts. Two were households in which all residents were under 20 years of age.

Based on knowledge of the sample cluster size of 2.18 adults per household, ($\frac{3649}{1675} = 2.18$) the number of adults can be estimated for the households where no household sheet was completed. These estimates appear below:

**Summary**

| | Frequency* | Estimate° | Total |
|---|---|---|---|
| Households | 1675 | | 1675 |
| Subjects | 3649 | $401(\frac{3649}{1675})$=873.58 | 4522.58 |
| Respondents | 3300 | | 3300 |
| Refusals | 195 | $148(\frac{3649}{1675})$=322.42 | 517.42 |
| Non-Respondents | 154 | $31(\frac{3649}{1675})$=67.53 | 221.53 |
| Other (N/S, No Answer, etc.) | | $222(\frac{3649}{1675})$=483.63 | 483.63 |

* Households for which there was $\geq$1 response

o Households in which there were no respondents

Therefore we estimate the following response rates:

| | |
|---|---|
| Households for which there was $\geq$1 response | $(\frac{3300}{3649})$=90.40% |
| Including all households where contact was made (i.e., excluding N/S, No Answer, but including complete household refusals and non-responses) | $(\frac{3300}{4038.95})$=81.70% |
| Including all households | $(\frac{3300}{4522.58})$=72.97% |

## 2.7 Linkage

The data having been collected, two stages of data linkage were carried out. Linkage refers to the joining of the survey data with data from another source. It was done via MCP numbers which were available for 2994 (or 90.7%) of the respondents. The remaining 306 respondents were those who refused to provide their MCP number or did not have one (foreign students or members of the security forces, for example). The data on the 2994 people were then linked with the data from two external sources.

The first source, termed 'hospital utilization' data, was extracted from computer tapes from the Department of Health and added to the survey data base.

These data provided the number of days a respondent spent in hospital (excluding hospitalizations due to pregnancy or delivery) for the four-year period from April 1981 to March 1985 and was the most up-to-date that could be obtained. The reason for the hospitalization was not used.

The second external source was termed 'physician consultation' data. This was obtained through computer records of doctors' insurance claims made to the Newfoundland Medical Care Commission. Due to the very strict confidentiality of this information, an *Order in Council* from the Provincial Cabinet was required before it could be released, a process which took approximately three months. Once released, it provided the number of physician consultations that a respondent had in the one year period corresponding very closely with calendar year 1985. Since diagnostic information was not made available, these were for all consultations, including pregnancy related visits. Again this was the most up-to-date information available.

**Summary**

| | Frequency |
|---|---|
| Number of Respondents | 3300 |
| Number Linked with MCP Data Files | 2994 |

**Summary of Those Linked With MCP Files**

| | Frequency |
|---|---|
| Number with $\geq$ 1 Doctor Visit over 1 year (including pregnancy) | 2434 |
| Number with $\geq$ 1 Hospital Day over 4 years (excluding pregnancy) | 599 |

## 2.8 Summary Suggestions for Future Telephone Surveys

The following are several suggestions which should be kept in mind when telephone surveys similar to this one are undertaken. This list is not intended to be exhaustive.

- In TELEPHONE SURVEY METHODS: Sampling, Selection and Supervision, Lavrakas (1987) suggests that selecting from the telephone directory is inadvisable if the proportion of non-coverage is estimated to be more than 10-15% and one is intending to generalize his results to the population at large. In this study the rate of non-coverage was estimated by the Newfoundland Telephone Company to be approximately 4%. If this proportion were to increase much beyond this point, random digit dialing (rdd) should be seriously considered since with rdd those with unlisted telephone numbers would be as likely to fall into the sample as those with listed numbers. This is important when a large proportion of households have unlisted numbers and people belonging to these households tend to exhibit certain characteristics. For example, according to Lavrakas, in the United States the most likely group of people to have unlisted numbers are lower income minority Americans.

- In estimating the number of telephone numbers required in our sampling pool to achieve a given number of completed questionnaires, the cluster size, estimated response rate and the estimated number of residential numbers in the section of interest in the telephone directory were considered. In addition to these, through a pilot study or possibly by contacting the telephone company, the number of 'not-in-service' numbers among the eligible households could have been estimated. An inflation factor might have also been used to compensate for other 'non-respondents', such as those whose numbers produced no answer

after seven calls and those who could not answer the questionnaire due to illness, for example. This might have eliminated the need to increase our sample pool size after the study had started. When this was increased it was done to compensate for the 'not-in-service' and 'no answer' numbers and for the decrease in the sample cluster size from the quoted census cluster size of 2.3 from Statistics Canada. Replacing such households if they were refusals could bias the results, but replacing them due to 'not-in-service' and 'no answer' numbers should not have this effect unless persons belonging to such households are not randomly distributed throughout the population. Before replacing these numbers in the future, it would be worthwhile to contact the telephone company for a breakdown of reasons for, and proportion of, 'not-in-service' lines.

- With respect to the field work it is strongly advised that the effort be made to directly supervise interviewers, particularly less experienced interviewers. In the event of space limitations, on each day one or two interviewers should be scheduled to conduct their interviews at the field office while the others carry theirs out at home. This should take place with as many interviews as possible at the beginning of the study with the frequency of supervised interviews dropping off as it progresses.

- For future questionnaires it would be advisable to break down the 'non-response' rate for each eligible member of the household into several categories, such as 'no response' after seven attempts, due to illness or due to absence during the survey period. A more thorough breakdown of reasons for 'non-response' could be useful when planning a similar type of survey in the future.

  In addition, all eligible households should have a household sheet completed even when no response is elicited from the unit so as to record whether this

was due to a not-in-service number or due to complete household refusal or non-response.

For complete household refusals, the attempt should be made to find out the number of eligible household members. This would increase the accuracy of the estimated number of refusals among these households. As it was, the number of refusals was estimated based on the sample cluster size from those 1675 households where the information was available.

Despite these practical problems, the survey was highly successful with a very low non-response rate.

# Chapter 3

# The Analysis of Data

The purpose of this study is to examine many socio-medical questions pertaining to people's lifestyles, health practices and utilization of health services. As such, information was obtained on some of the many variables associated with these aspects of people's lives.

The data were collected and first briefly explored by looking at frequency distributions and descriptive statistics. As is the case in many studies in the social sciences, the data collected in the present survey were, for the most part, categorical. This chapter, therefore, will deal with analytical tools for categorical data. Since categorical data are often presented as cross-tabulations, we will look at two-way and multi-way tables. Tests of hypotheses of independence will be considered as will several of the many measures of association developed for just such analysis of categorical data. Strengths and weaknesses of these measures will be discussed. As the emphasis is on the application as opposed to the mathematical development of these measures of association, they are not rigorously dealt with from the mathematical point of view. After this preliminary analysis we will further examine the manner in which variables interact with one another. To this end we generate models for given sets of variables. To do so we employ such statistical techniques as logistic regression

and loglinear analysis.

Many interesting questions existed for the research team so that during the analysis many different variables were explored. From the perspective of the research team and from the socio-medical point of view, all those explored are of interest. It is not the primary purpose of this report, however, to present medical findings. For this reason, only a small number of the variables are focused upon since to do otherwise would result in much repetition in this chapter. This subset of variables will suffice as illustrations throughout the remainder of this report and will be discussed in varying detail at the time of illustration.

SPSS-X and BMDP were the primary statistical packages used in the analysis. Minitab was also used to a lesser extent. All analyses were done on the VAX Cluster running VMS in the Department of Computing Services at Memorial University of Newfoundland.

## 3.1 Contingency Table Analysis

A contingency table classifies data according to some categorical criterion. We may have an $r \times c$ contingency table, for example, which crosses $r$ levels of variable $A$ with $c$ levels of variable $B$. Our data are classified according to the particular category of $A$ and $B$ to which they belong. The categories of a given variable are mutually exclusive and any given person or item can fall into one and only one cell of the the contingency table.

In examining our data in this study, we wanted to see if two variables in our table were independent, and if not independent, to what degree they could be considered related or associated. Our hypothesis is given as:

$H_o$: There is no association, versus

$H_1$: The null hypothesis is not true

In the sections which follow, we shall test this hypothesis and discuss, in general, the measures of association which may be used to examine to what degree the variables may be related to one another.

### 3.1.1 Tests of Independence in Two-Way Tables

In studies such as this one, a simple random sample (of households in this instance) is taken and only the sample size $n$ is fixed. A variety of questions are asked of those in the sample. This is as opposed to the instance when marginal totals are fixed. This would be the case, for example, if prior to the study one were to fix the number of males and females to interview. No marginal totals were fixed in our study.

We consider a two-way ($r \times c$) cross-tabulation of two discrete categorical variables, $A$ and $B$, where $f_{ij}$ is the frequency of observations in the cell of the contingency table corresponding to row $i$ and column $j$ – that is, corresponding to levels $i$ and $j$ of variables $A$ an $B$, respectively. The marginal row frequency $f_{i.} = \sum_{j=1}^{c} f_{ij}$ is the sum of the frequencies of level $i$ of variable $A$ over all levels of variable $B$. Similarly, the marginal column frequency is $f_{.j} = \sum_{i=1}^{r} f_{ij}$. The total frequency of all subjects is given by $f_{..} = n$. Expressed in terms of observed proportions, $p_{ij}$ is the observed proportion in row $i$ and column $j$. $P(A = i) = P_{i.} = \sum_{j=1}^{c} P_{ij}$ and $P(B = j) = P_{.j} = \sum_{i=1}^{r} P_{ij}$. Under the assumption of independence of $A$ and $B$, $P(A = i, B = j) = P_{i.}P_{.j} = P_{ij}$ where $P_{i.}$ and $P_{.j}$ are the marginal probabilities and $P_{ij}$ is the joint probability.

In what follows, $f_{ij}$ and $p_{ij}$ will denote observed frequencies and proportions, respectively while $\hat{F}_{ij}$ and $\hat{P}_{ij}$ will denote the corresponding estimated expected val-

ues. The standard maximum likelihood estimates of the marginal probabilities are $\hat{P}_{i.} = f_{i.}/n$ and $\hat{P}_{.j} = f_{.j}/n$.

We may test the independence of $A$ and $B$ by looking at the $\chi^2$ test statistic, $X^2$, which is commonly used to test for homogeneity or independence of variables. As is well known, under our null hypothesis of independence, $X^2$ has an approximate $\chi^2_{(r-1)(c-1)}$ distribution where

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - \hat{F}_{ij})^2}{\hat{F}_{ij}}$$

The lower bound on this statistic is 0, which is achieved when $f_{ij} = \hat{F}_{ij}$ for all $i, j$. Provided there are no zero marginal totals, the upper bound is $n(q-1)$ where $q$ =min{$r$,$c$}. Cramér (1946) states that any row or column consisting entirely of zeros may be discarded and Blalock (1972) shows how, under this assumption of non-zero marginal frequencies, the upper bound on $X^2$ is $n(q-1)$. Without this restriction there is no upper bound.

Although $X^2$ is easy to calculate and apply and is frequently used, it should be used with caution when the sample size is large, as is the case here. Being sensitive to sample size, the test statistic will grow as $n$ increases and hence the null hypothesis may well be rejected simply because $n$ is large, rather than because the hypothesis is not true. In discussing $X^2$, Reynolds (1977a), for example, comments that "one can always find a significant relationship by making the sample large enough. In public opinion surveys, where $n$ often exceeds 1500, the difficulty of separating substantive from statistical significance is particularly acute." Also recall from our section on sample size that we must be careful in our reliance on $X^2$ if our table contains cells with zero frequency. As mentioned in that section, it is generally suggested that this test statistic be used only if there are no cells with zero frequency and a minimum of 80% of the cells have 5 or more observations.

**Yates' $\chi^2$ corrected for continuity: $X_c^2$**

$X^2$ is known as the Pearson $\chi^2$ test statistic. Theoretically it is appropriate only when the expected values in the contingency table are large, as only then can it be assumed to have an approximate $\chi^2$ distribution. Therefore the suitability – or indeed the validity – of this test statistic may be questionable when these values are small. Yates suggested a factor to correct for this situation in a 2×2 table. We will denote Yates' corrected $X^2$ by $X_c^2$ where

$$X_c^2 = \frac{n(|f_{11}f_{22} - f_{12}f_{21}| - \frac{n}{2})^2}{f_{1.}f_{2.}f_{.1}f_{.2}}$$

Maxwell (1961, 1978) is a proponent of $X_c^2$ claiming that it should be favoured over $X^2$ even if the expected values are at least 5 and that, in any event, it must be used when the sample size is small. Everitt (1977) also recommends it while pointing out that there has been debate regarding its use in all cases. Fingleton (1984) avoids using it in his discussion, citing Fienberg. And Fienberg (1977) suggests that the use of $X_c^2$ may not be appropriate if the reason for using it is to correct $X^2$ so that it more closely approximates a $\chi^2$ distribution when the sample size is large. He, like Grizzle (1967) and Conover (1974) before him, warns that $X_c^2$ may lead to a test that is too conservative; the null hypothesis is not rejected as frequently as it should be in 2×2 tables. There are many contributions to the literature which debate the merits of the continuity correction $X_c^2$ over $X^2$.

**Fisher's exact test**

Another alternative to $X^2$ for 2×2 tables is Fisher's exact test (Everitt 1977, Reynolds 1977a, Upton 1978) which is given by

$$P = \frac{f_{1.}!f_{2.}!f_{.1}!f_{.2}!}{f_{11}!f_{12}!f_{21}!f_{22}!f_{..}!}$$

Rather than approximating a $\chi^2$ distribution, this calculates the exact probabilities.

As with $X_c^2$, this may be used when the expected cell frequencies are small. This test statistic may be recommended when the sampling scheme involves fixing marginal totals. Fisher's exact test is a one-tailed test as opposed to the two-tailed $X^2$ and $X_c^2$ tests. In tables with large values for cells and for row and column marginals, this is cumbersome to calculate. For 2×2 tables, statistical packages such as SPSS-X and BMDP calculate $P$ only when the minimum expected cell frequency is less than 20; if this frequency is at least 20, then test statistics which have approximate $\chi^2$ distributions are substituted.

When examining fourfold tables in our study, we do not require such alternatives to the $\chi^2$ test statistic since with our large sample size and our variables under consideration, we should not have cells with such small expected cell frequencies as would warrant these alternatives.

Likelihood-Ratio Test: $G^2$

The likelihood-ratio test, $G^2$, is also used to test for independence. Again, if the expected cell frequencies are large, it approximates a $\chi^2_{(r-1)(c-1)}$ distribution. It is given by

$$G^2 = 2\sum_{j=1}^{c}\sum_{i=1}^{r} f_{ij}[\log(\frac{f_{ij}}{\hat{F}_{ij}})]$$

where log is the natural logarithm. $G^2$, like $X^2$, should be used with caution if at all, when expected cell frequencies are small. We are not, as a general rule, seriously affected by this in our study particularly in lower dimensional tables.

It has been known for some time that smoking adversely effects one's health. Given the amount of public awareness of and concern about the effects of smoking on health, the research team was interested in studying the relationship between smoking habits of the general public and their self-assessed health status Self-assessed health status is a measure of health that has been proposed in the social medical literature

as a valid substitute for the very costly evaluation of health by a medical team.

Respondents were asked questions pertaining to smoking habits, and from their answers a variable was constructed which categorized each respondent as having never smoked, as a former smoker (having given up smoking for at least one year), or as a current smoker. The respondents were also asked to rate their health as poor, fair, good, or excellent. From these two variables we consider the 3×4 table below, where the values in parenthesis are the expected values.

Table 3.1

| | Health Status | | | | |
|---|---|---|---|---|---|
| Smoke | poor | fair | good | excellent | Totals |
| never smoked | 10 | 172 | 696 | 414 | 1292 |
| | (20.8) | (214.6) | (696.9) | (359.8) | |
| former smoker | 14 | 125 | 428 | 250 | 817 |
| | (13.1) | (135.7) | (440.7) | (227.5) | |
| current smoker | 29 | 251 | 656 | 255 | 1191 |
| | (19.1) | (197.8) | (642.4) | (331.7) | |
| Totals | 53 | 548 | 1780 | 919 | 3300 |

$G^2 = 64.322$
p= .0000

$X^2 = 63.087$
p= .0000

(df= 6)

The observed significance level, or p-value, which we denote by p, is the probability of getting a test statistic value at least as extreme as the value observed. Here we reject the hypothesis of independence between smoking habit and self-assessed health habits; these two variables appear to be related in some way. With the $X^2$ and $G^2$ statistics we cannot assume causality although, from a medical perspective, one would probably surmise that if dependence is indicated then it is more likely that self-assessed health status is dependent upon smoking habit than the reverse.

In this particular example there are no cells with a frequency less than 5, but the sample size is quite large and it could be that our test statistics were large enough to cause us to reject our hypothesis not because the variables are truly independent but because $X^2$ and $G^2$ are sensitive to the large sample size. Because of this, with large

sample sizes we should not rely entirely on values of $\chi^2$ statistics. In a later section we will discuss statistics which try to compensate for this and will also consider measures of association which may shed more light on the relationship which exists between variables which are apparently not independent.

### 3.1.2 Partitioning $\chi^2$ Test Statistics in Two-Way Tables

Often we are not interested only in the hypothesis of independence between variables in a contingency table but also in subhypotheses within this table. For a medical researcher this is the case with the hypothesis we have just explored. It was an important table and further analysis was attempted by examining subhypotheses through partitioning.

There are methods for partitioning tables which enable one to divide the original table into subtables on which subhypotheses may be subsequently tested using a $\chi^2$ test statistic such as Pearson's $X^2$ or the likelihood-ratio $G^2$. Although different methods exist for doing so, we shall only give an example using the method used by Goodman (1968) (see Reynolds 1977a or Agresti 1984, for example). As pointed out many times in the literature, a $\chi^2$ statistic can be decomposed into component parts such that the degrees of freedom of the overall statistic is equal to the sum of the degrees of freedom of those parts. In an $r \times c$ table, for example, we can partition our overall $\chi^2$ into as many as $(r-1)(c-1)$ component parts since there are that many degrees of freedom. In this case, each component part would correspond to a $2 \times 2$ table each which would be tested for independence with a $\chi^2$ test statistic with 1 degree of freedom. Pearson's $X^2$ has been used with such partitions; however we use $G^2$ since when partitioned the component parts of $X^2$ sum approximately to the $X^2$ of the original table whereas the component parts of $G^2$ sum exactly to the overall $G^2$.

Let us look again at our table of *smoking×health status.* In that original table we rejected our hypothesis of independence between these two variables. Prior to examining that table we were interested in the independence of these variables with smoking as a dichotomy – either one smokes or does not. With this variable still dichotomous, we were also interested in the independence of the two variables when an individual assesses his health status as either poor or fair, or as good or excellent. To this end, let us re-examine our 3×4 table, applying a method of partitioning given by Goodman as stated in Reynolds (1977a). Under this method we partition the original 3×4 table into two parts. One subtable consists of the first 2 rows and all 4 columns to give us a 2×4 table (3.1a). That is, we drop the current smokers from our table. Our second subtable is also a 2×4 table (3.1b) where one of the rows will be the row ignored in the first subtable (the current smokers) and the other row is the sum of the rows used in that first subtable, namely the former smokers and those who never smoked. Let us look at the first subtable of our partition.

Table 3.1a

| | Health Status | | | | | |
|---|---|---|---|---|---|---|
| Smoke | poor | fair | good | excellent | Totals | |
| never smoked | 10 | 172 | 696 | 414 | 1292 | $G^2 = 5.681$ |
| | (14.7) | (181.9) | (688.6) | (406.8) | | p= .1282 |
| former smoker | 14 | 125 | 428 | 250 | 817 | $X^2 = 5.824$ |
| | (9.3) | (115.1) | (435.4) | (257.2) | | p= .1205 |
| Totals | 24 | 297 | 1124 | 664 | 2109 | (df= 3) |

Noting the values for the test statistics for this table, we say that they are not significant and therefore we do not reject the subhypothesis of independence of the two variables when current smokers are not considered. This is a rather interesting finding as it implies that those who have given up smoking for at least one year do not appear to rate their health status differently than those who never smoked.

Now let us look at table 3.1b, the second partition of our 3×4 table.

Table 3.1b

| Smoke | Health Status | | | | Totals |
|---|---|---|---|---|---|
| | poor | fair | good | excellent | |
| not current (former/never) | 24 (33.9) | 297 (350.2) | 1124 (1137.6) | 664 (587.3) | 2109 |
| current smoker | 29 (19.1) | 251 (197.8) | 656 (642.4) | 255 (331.7) | 1191 |
| Totals | 53 | 548 | 1780 | 919 | 3300 |

$G^2 = 58.641$ p= .0000

$X^2 = 58.567$ p= .0000 (df= 3)

This is significant, so we reject the subhypothesis of independence of the two variables when smoking is dichotomized as current and not current smokers. Those who do not smoke currently – whether they have never smoked or are former smokers – appear to rate their health status differently than those who are current smokers.

Recall that prior to examining the original table we were also interested in the independence of the two variables when an individual rates his health status as either poor or fair, or as good or excellent. Continuing to partition table 3.1b, we consider the tables which follow. In each case the smoking variable is dichotomized as in table 3.1b. In the first subtable, 3.1c, we only look at those people with poor or fair self-assessed health status.

Table 3.1c

| Smoke | Health Status | | Totals |
|---|---|---|---|
| | poor | fair | |
| not current | 24 (28.3) | 297 (292.7) | 321 |
| current smoker | 29 (24.7) | 251 (255.3) | 280 |
| Totals | 53 | 548 | 601 |

$G^2 = 1.539$ p= .2148

$X^2 = 1.543$ p= .2141 (df= 1)

This is not significant, hence we do not reject the subhypothesis of independence of the two variables as they stand here. It is interesting that for those who rate their health as less than good, the fact that they are current smokers or not current smokers is independent of whether they rate their health as either poor or fair.

In our next subtable, 3.1d, we consider the remainder of our respondents, namely those who rated their health as good or excellent.

Table 3.1d

| Smoke | Health Status: good | Health Status: excellent | Totals |
|---|---|---|---|
| not current | 1124 | 664 | 1788 |
| | (1179.2) | (608.8) | |
| current smoker | 656 | 255 | 911 |
| | (600.8) | (310.2) | |
| Totals | 1780 | 919 | 2699 |

$G^2 = 22.866$ p= .0000
$X^2 = 22.477$ p= .0000
(df= 1)

As this is significant we reject the subhypothesis of independence of the variables when only those with good or excellent self-assessed health status are considered. For this sub-group, those who do and do not currently smoke appear to rate their health status differently.

Finally we examine more closely another subtable (3.1e) in which we were particularly interested and which prompted the second stage of partitioning. In this instance, with all respondents included, the self-assessed health status variable is coded as either poor or fair, or as good or excellent.

Table 3.1e

| Smoke | Health Status: poor/fair | Health Status: good/excellent | Totals |
|---|---|---|---|
| not current | 321 | 1788 | 2109 |
| | (384.1) | (1724.9) | |
| current smoker | 280 | 911 | 1191 |
| | (216.9) | (974.1) | |
| Totals | 601 | 2699 | 3300 |

$G^2 = 34.236$ p= .0000
$X^2 = 35.111$ p= .0000
(df= 1)

Since this is significant, we again reject the subhypothesis of independence of the two variables when they are both dichotomized as seen in the subtable. Those who do not currently smoke rate their health differently from those who do smoke. The non-current smokers are more inclined than the current smokers to rate their health

as good or excellent rather than poor or fair. This confirms previous work which has acknowledged for some time that smoking has detrimental effects on health. Although we cannot assume causality here, we can state that using this dichotomy a person's smoking status is not independent of his self-assessed health rating.

Note that the component subtables of table 3.1b, namely tables 3.1c, 3.1d and 3.1e, give $G^2$ values which sum exactly to the $G^2$ value for table 3.1b. In this partitioning, the $G^2$ associated with table 3.1c contributes much less to the $G^2$ of table 3.1b than does the $G^2$ of table 3.1d or 3.1e. The contributions of the $G^2$ and $X^2$ statistics obtained from the subtables of the original table are summarized below:

Table 3.1f

| Table | /Subtable | df | $G^2$ | $X^2$ | p ($G^2$) | p ($X^2$) |
|---|---|---|---|---|---|---|
| | Initial Partitioning of Table 3.1 | | | | | |
| 3.1 | original | 6 | 64.322 | 63.087 | .0000 | .0000 |
| 3.1a | never vs former smokers on assessing health as poor, fair, good or excellent | 3 | 5.681 | 5.824 | .1282 | .1205 |
| 3.1b | not current vs current smokers on assessing health as poor, fair, good or excellent | 3 | 58.641 | 58.567 | .0000 | .0000 |
| | Further Partitioning of Table 3.1b | | | | | |
| 3.1c | not current vs current smokers on assessing health as health as poor or fair | 1 | 1.539 | 1.543 | .2148 | .2141 |
| 3.1d | not current vs current smokers on assessing health as good or excellent | 1 | 22.866 | 22.477 | .0000 | .0000 |
| 3.1e | not current vs current smokers on assessing health as poor/fair or good/excellent | 1 | 34.236 | 35.111 | .0000 | .0000 |

The method used by Goodman can be further extended so that any $r \times c$ table can be partitioned into $(r-1)(c-1)$ 2×2 tables. For a nice illustration on how to do this, see Reynolds (1977a).

For additional discussions on methods of partitioning $\chi^2$ statistics, see Lancaster (1949), Kimball (1954), or Maxwell (1961). In the method used in the above partitioning, subhypotheses of interest are decided upon prior to the testing of the overall table. Some authors comment that the subhypotheses can be suggested once the original table has been examined. From the original table the researcher can focus on those cells which contribute the most to the overall statistic. Based on these cells, the researcher may then decide upon which subtables he wishes to examine. Specific guidelines for this partitioning are given by Upton (1978), Iverson (1979), and Freeman (1987), for example.

As convenient and attractive as Goodman's method is, one should be careful when examining particular subtables based on decisions made *after* studying the original table as this is contrary to the underlying assumption of randomness. Maxwell (1961) and Everitt (1977), for example, warn against this. It is advisable that decisions be made *a priori* if the intention is to draw conclusions from the test of hypotheses. When one has no idea in advance which subtables might be of interest, the only option may be to choose subtables after examination of the original table. However, conclusions should not be drawn in this instance. Rather the investigator might use this as a means of exploratory analysis and any findings may suggest possible subtables to investigate in future studies.

### 3.1.3 Measures of Association

By rejecting a hypothesis of independence between smoking and health status, we are claiming that some association exists between the two variables. This section looks at how we might judge to what extent they and other variables are associated.

With the $X^2$ statistic, when the observed and expected values are equivalent

and hence the value of $X^2$ is 0, there is no association. All else being equal, the larger $X^2$, the greater the association. We must be cautious about relying on this as a measure of association just as we must when using it as a test statistic. Since the value of $X^2$ increases as the sample size increases, a large value for $X^2$ may simply reflect a large sample size rather than a strong association.

Pearson proposed correcting for this by dividing by $n$ to get a measure of association between the two variables. Other measures have also been proposed which are more or less appropriate depending upon the nature of the data. To see how much of a relationship there is, we turn to measures of association formulated specifically for categorical data. Which measures are used will depend on the data and the variables of interest to the researcher. The variables, for example, may be nominal or ordinal. Or rather than looking at the association between two variables, we may be interested in the level of agreement between spouses, say, as they consider the same question.

We will first look at measures of association which are based on Pearson's $X^2$ and then at the cross-product ratio and measures based on this ratio. Following this, we will look at measures of proportional reduction in predictive error and of agreement. Finally we will focus on those which take account of the ordinality of variables.

Several measures of association are described of which a small subset are used in the analysis. These were deemed to be the most appropriate and useful to our health survey data. The others, although not employed in this study, are described because the measures more commonly used and discussed in the literature should be mentioned briefly in a report which deals with the analysis of categorical data so that one may ascertain why certain measures of association were considered to be more suitable than others. In any analysis of categorical data, while not all measures of

association are appropriate, one need not limit himself to one and only one 'correct' measurement.

## Measures Based on $X^2$

$\chi^2$ test statistics were discussed as they pertained to the testing of independence of categorical data. Now we shall look at several measures of association based on Pearson's $\chi^2$ test statistic, $X^2$.

### Phi-Squared, $\phi^2$, The Mean Square Contingency Coefficient

Recall that $\chi^2$ is sensitive to the sample size, $n$, in as much as its magnitude is proportional to $n$. As a result, it cannot be considered a reliable measure of association. Pearson removed this sample size effect with the measure of association, $\phi^2$ which is estimated as $X^2$ divided by $n$. That is,

$$\phi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(P_{ij} - P_{i.}P_{.j})^2}{P_{i.}P_{.j}} \qquad \text{estimated by} \qquad \hat{\phi}^2 = \frac{X^2}{n}$$

Since $0 \leq X^2 \leq n(q-1)$, $q = \min\{r,c\}$ in an $r \times c$ contingency table, it follows that $0 \leq \phi^2 \leq q-1$. This is assuming that there are no zero marginal totals. Without this assumption the upper limit of $\phi^2$ is infinity. The minimum value of zero is achieved when there is no association between the variables. The maximum value of $q-1$ is attained when there is strict perfect association in a square table or implicit perfect association in a non-square table. Strict perfect association, which can only occur in a square table, is attained when each row *and* each column has one non-zero entry. Implicit perfect association in a non-square $r \times c$ table means that for each row *or* column (but clearly not both) there is only one non-zero cell. And all else being equal, the closer the value to $q-1$, the stronger the association. Although $\phi^2$ is not sensitive to sample size it is clear that it is still dependent on the dimensions of the

contingency table. This can make interpretation of $\phi^2$ difficult if it does not take a value of 0 or $q-1$. Another disadvantage of $\phi^2$ is that it is sensitive to the marginal distributions of the variables. The greater the marginal variation in the variables the further is the value of $\phi^2$ from its upper bound. Hence this is not a good measure of association if the marginals are highly skewed. In our study, when cross-tabulating health practice variables by health status and hospital utilization variables, quite often one or both marginal distributions from a table were skewed, so although an appropriate measure occasionally, it was not one of the more favourable ones for our study.

Since $\phi^2$ is not sensitive to $n$, it may be used to compare tables provided that the marginal distributions are not highly skewed and are similar between tables. The medical researcher might be interested in comparing tables from the study with one another or with those from a similar study. If the marginal distributions for the tables are alike for the samples from each study, for instance, we might consider this an acceptable comparative measure of association. If we wish to compare tables but the marginals are not the same from table to table, it would not then be advisable to use this as a measure. When comparing tables or when the marginals are highly skewed, in addition to displaying the original table Reynolds (1977a) suggests standardizing tables. Garson (1976) discusses the maximum value attainable for $\phi^2$ when the marginals for the two variables differ, where this maximum value depends upon the marginals.

In the special case of the 2×2 table, $\phi^2$ reduces to

$$\phi^2 = \frac{(P_{11}P_{22} - P_{12}P_{21})^2}{P_{1.}P_{2.}P_{.1}P_{.2}} \qquad \text{estimated by} \qquad \hat{\phi}^2 = \frac{(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1.}f_{2.}f_{.1}f_{.2}}$$

which is the same as the square of Pearson's correlation coefficient $\rho$, which is discussed in most elementary statistic texts. $\rho$ can be considered a measure of association

or correlation between two variables with $\rho^2$ being the percent of variation of the dependent variable explained by the independent variable. Of course, measures based on $X^2$ do not assume that one variable is dependent and the other independent. That is, they are symmetric measures.

### Pearson's Contingency Coefficient $C$

To overcome the fact that $\phi^2$ can exceed 1, Pearson introduced the measure of association, $C$ where

$$C = \sqrt{\frac{\phi^2}{\phi^2 + 1}} \qquad \text{estimated by} \qquad \hat{C} = \sqrt{\frac{\hat{\phi}^2}{\hat{\phi}^2 + 1}}$$

which can clearly never exceed unity. This measure is theoretically bounded by 0 and 1, taking the value of 0 when the variables are independent.

The upper limit of 1 cannot be attained in practice. The maximum value of $C$ which can be attained is $\sqrt{\frac{q-1}{q}}$ where $q = \min\{r, c\}$ and this occurs under strict or implicit perfect association in a square or a non-square table, respectively. So we see that at its maximum value under perfect association, the value of $C$ relies upon the number of rows and columns, approaching unity as the number of rows and columns increases. Garson (1976) comments that for this reason some social scientists suggest using this measure only when tables are at least 5×5. Even under strict perfect association unity will not be reached. Since this can make interpretation difficult, Reynolds (1977a), for example, suggests dividing $C$ by the maximum value of $C$ in a square contingency table. Given that our tables are usually not as large as 5×5 nor square, this was not generally an appropriate measures to use in our study.

### Tschuprow's $T$

Because even under strict or implicit perfect association the maximum value of $C$ depends upon the number of rows and columns, Tschuprow proposed another measure of association, $T$, where

$$T = \left[\frac{\phi^2}{[(r-1)(c-1)]^{\frac{1}{2}}}\right]^{\frac{1}{2}}$$

Note that when $r = c = 2$, $T = \phi$.

Recall that $0 \leq \phi^2 \leq q - 1$ where $q = \min\{r, c\}$. So again $T$ will be zero when the variables are independent. The maximum value which can be attained is given by

$$\left[\frac{\min(r-1, c-1)}{[(r-1)(c-1)]^{\frac{1}{2}}}\right]^{\frac{1}{2}} = \left[\frac{\min(r-1, c-1)}{\max(r-1, c-1)}\right]^{\frac{1}{4}}$$

Unlike $C$ which *approaches* 1, this measure of association will *attain* 1 under strict perfect association. That is, it will achieve unity when there is perfect association in an $r \times r$ table regardless of how large the table is. (Recall Garson's advice regarding the use of $C$ only when tables are at least 5×5.) When the table is not square this is not true and $T < 1$. Liebetrau (1983) warns that the maximum value of $T$ becomes quite small if $r$ and $c$ are not almost equal. Even though there were some square tables in our health study, most were not and the dimensions were not always close to square. So although an improvement over $C$, this again was not a measure of association useful for our analysis.

### Cramér's $V$

Yet another measure of association based on $X^2$ is Cramér's $V$. This quantity, introduced by Cramér in 1946, is a standardized $\phi$ given by

$$V = \left(\frac{\phi^2}{q-1}\right)^{\frac{1}{2}} = \left(\frac{\phi^2}{\min(r-1, c-1)}\right)^{\frac{1}{2}}$$

where $0 \leq V \leq 1$.

Notice that when $r = c = 2$, $V = T = \phi = \rho$ and in a $2 \times c$ table, $V = \phi$. Furthermore, when we have any $r \times c$ table,

$$V = \left( \frac{\phi^2}{\min(r-1, c-1)} \right)^{\frac{1}{2}} \geq \left( \frac{\phi^2}{[(r-1)(c-1)]^{\frac{1}{2}}} \right)^{\frac{1}{2}} = T$$

with equality holding only when $r = c$.

This measure $V$ is preferable to $T$ since it can achieve unity for all $r \times c$ tables in the case of strict or implicit perfect association. In other words, although $T$ may attain its maximum value in square tables, $V$ may attain this value for any $r \times c$ table. Agresti (1984) suggests using measures of association such as $V$ for comparing the degree of relationship between tables as opposed to using it as an association measure for a given table but warns against this if the marginal distributions of the tables being compared are not similar.

For all the aforementioned measures of association based on $X^2$, confidence limits may be obtained if one calculates the asymptotic variance. See Bishop, Fienberg and Holland (1975), Kendall and Stuart (1979), or Liebetrau (1983).

Despite their ease of calculation, these are not necessarily the best measures of association to employ. As pointed out, they tend to be sensitive to the marginal distributions of the variables and to the table dimensions and so unless the variables are independent or perfectly associated they are difficult to interpret. They can sometimes be useful for comparing tables of the same dimensions but caution is extended here as well if the marginal distributions differ very much from table to table. Since our cross-tabulations of health variables have marginal distributions which are frequently skewed these measures of association were infrequently calculated. Other measures of association are more appropriate for our health data. Nonetheless there was occasion when we wished to compare tables with similar marginal distributions

and given the pros and cons of the aforementioned, Cramér's $V$ was used for this comparison. Examples are not presented here as such tables are discussed in the later section on loglinear analysis.

In what follows we will look at alternative measures of association available for $r \times c$ contingency tables and will begin by exploring those intended specifically for 2×2 tables.

## The Cross-Product Ratio and Measures Based on it

### Cross-Product or Odds Ratio, $\alpha$

A frequently used measure of association is the cross-product or odds ratio. It has some excellent features and is useful in aiding in the understanding of loglinear analysis since it plays an important part in the development of loglinear models. Because loglinear models can be an important tool in the analysis of health data like we have in this current project, this cross-product ratio is discussed in some detail. First we shall look at it as it pertains to fourfold tables after which we shall look at two measures based on the cross-product ratio.

Consider the fourfold table where $f_{ij}$ is the observed frequency in the cell corresponding to row $i$ and column $j$.

| | B | |
|---|---|---|
| A | $B_1$ | $B_2$ |
| $A_1$ | $f_{11}$ | $f_{12}$ |
| $A_2$ | $f_{21}$ | $f_{22}$ |

If our two variables, $A$ and $B$, are independent we would conclude that knowing a person's characteristic on one variable will not enlighten us regarding which category of the other variable he belongs to. Referring to our table we say that conditional

upon belonging to category 1 of variable $A$ (that is, $A_1$), the odds of belonging to category 1 (as opposed to category 2) of variable $B$ are $f_{11}/f_{12}$. Similarly, given that one belongs to $A_2$, the odds of belonging to category $B_1$ (as opposed to $B_2$) are $f_{21}/f_{22}$. If these two ratios are equivalent, knowing whether a person has characteristic $A_1$ or $A_2$ does not help us identify whether he will have characteristic $B_1$ or $B_2$; that is, $A$ and $B$ are independent. We can think of this equivalently in terms of the cross-product or odds ratio, $\alpha$. This is simply the ratio of the two aforementioned ratios so that,

$$\alpha = \frac{P_{11}/P_{12}}{P_{21}/P_{22}} = \frac{P_{11}P_{22}}{P_{12}P_{21}} \quad \text{which is estimated by} \quad \hat{\alpha} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{f_{11}f_{22}}{f_{12}f_{21}}$$

Clearly, if the two odds are equal then the ratio of the odds, $\alpha$ , is unity. Hence we say that if variables $A$ and $B$ are independent or not associated, then $\alpha = 1$. This is clear when expressed in terms of the ratio of two equal odds. But what if the variables are associated? The range of $\alpha$ is $0 \leq \alpha < \infty$ with the lower bound being achieved when $P_{11}$ and/or $P_{22}$ is zero and the upper bound being attained when $P_{12}$ and/or $P_{21}$ is zero. In other words, the upper and lower bounds may be attained under either strict perfect (opposite cells off a diagonal are both zero) or under weak perfect (only one cell is zero) association. As Reynolds (1977a) points out, that it can achieve its upper or its lower bound under weak perfect association may be considered a weakness by some. Agresti (1984) refers to Gart and Zweifel (1967) in mentioning that if one has no reason in theory for suspecting $P_{ij} = 0$ for any $i, j$ then the estimator $\hat{\alpha}^* = \frac{(f_{11}+.5)(f_{22}+.5)}{(f_{12}+.5)(f_{21}+.5)}$ might be used in lieu of $\hat{\alpha}$. This is also recommended by Upton (1978).

One feature of the cross-product ratio is that under row *or* column interchange only, the direction of the association changes while the magnitude remains the same. If we denote the original cross-product ratio as $\alpha$ and the cross-product ratio resulting

from row or column interchange as $\alpha'$, then $\alpha = 1/\alpha'$. When $0 \leq \alpha < 1$ the association is 'negative' and when $1 < \alpha < \infty$ it is 'positive'. This is clearer if we express the odds ratio as a log odds ratio. That is, take the natural logarithm of $\alpha$ to give

$$\log \alpha = \log \frac{P_{11}P_{22}}{P_{12}P_{21}}$$

The measure of association, $\log \alpha$, now has a range of $-\infty < \log \alpha < \infty$ with no association existing when $\alpha = 1$ or $\log \alpha = 0$. Now we see that if the two ratios $\alpha$ and $\alpha'$ are such that $\alpha = 1/\alpha'$, then $\log \alpha = \log \frac{1}{\alpha'}$ or $\log \alpha = -\log \alpha'$ so clearly the ratios have the same magnitude of association but in opposite directions.

Given public concern about diet, the research team was interested in the eating habits of the general population. To find out a little about their eating habits, one question people were asked was if they made any conscious effort to limit the amount of red meat in their diet for health reasons. The medical researcher was interested in whether males and females respond differently to this question. The table below cross-tabulates sex with whether or not one limits the amount of red meat in his or her diet.

| | Limit Red Meat | |
|---|---|---|
| Sex | No | Yes |
| Male | 1097 | 416 |
| Female | 1024 | 763 |

If we were to calculate a $\chi^2$ test statistic here we would reject a hypothesis of independence of these variables, but let us look more closely at this table using the cross-product ratio. If gender is thought of as being fixed, then conditional on being male, the odds of not limiting red meat (as opposed to limiting it) are 1097 to 416, or $\frac{1097}{416} = 2.64$, so that men appear *not* to limit red meat in their diets more than two and a half times as frequently as they *do*. Given that one is female, on the other hand, the odds of not limiting red meat (as opposed to limiting it) are 1024 to 763,

or $\frac{1024}{763} = 1.34$ so that women do *not* limit red meat less than one and a half times as frequently as they *do*. The ratio of these odds is

$$\hat{\alpha} = \frac{1097/416}{1024/763} = 1.96 \qquad \text{so that} \qquad \log \hat{\alpha} = .68.$$

Men appear less likely than women to limit the red meat in their diet, in a ratio of approximately 2 to 1. So then, here the odds ratio is a way of measuring the strength of association between a person's sex and his or her limitation on eating red meat; it gives a clear picture of how these variables are related.

We may wish to have a confidence interval for this measure or may wish to compute a statistic to test its significance. We can calculate this, as for large samples there exist estimates of the variance of this measure of association (see Bishop et al. 1975, or Fienberg 1977).

$$\hat{\sigma}^2_{(\hat{\alpha})} = \hat{\alpha}^2 \left( \frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}} \right) \quad \text{and} \quad \hat{\sigma}^2_{(\log \hat{\alpha})} = \frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}}$$

provided $f_{ij} > 0$ for all $i, j$. If any $f_{ij} = 0$, one simple alternative is to add $\frac{1}{2}$ to each observed cell when computing $\alpha$ or $\log \alpha$ (see. for example, Reynolds 1977a or Liebetrau 1983).

For large samples, $\hat{\alpha}$ and $\log \hat{\alpha}$ are approximately normally distributed with means $\alpha$ and $\log \alpha$, respectively. For large $n$, the approximate $100(1-p)\%$ confidence interval for $\log \alpha$ is

$$\log \hat{\alpha} \pm Z_{p/2}(\hat{\sigma}_{(\log \hat{\alpha})})$$

Since for large samples

$$X^2 = \frac{(\log \hat{\alpha})^2}{\hat{\sigma}^2_{(\log \hat{\alpha})}}$$

is asymptotically distributed as $\chi^2_1$, the statistic $X^2$ may be used to test for independence of the two variables in a fourfold table. Expressing $X^2$ in this way, we

calculate

$$X^2 = \frac{\left[\log \frac{(1097)(763)}{(416)(1024)}\right]^2}{\frac{1}{1097} + \frac{1}{416} + \frac{1}{1024} + \frac{1}{763}} = 81.43$$

On the basis of this test statistic we would reject our hypothesis of independence of the two variables.

Aside from its relative ease of interpretation, the odds ratio has a couple of other important properties. First, it is invariant under row *and* column interchange and, as already mentioned, under row *or* column interchange, only the direction of the association changes. If we change the rows only of our general fourfold table, we get the reciprocal of our original cross-product ratio. If the columns of this table are also interchanged then we get back our original table. So our measure of association is symmetric; it does not matter which of our variables we consider the dependent or independent variable.

Another feature of the cross-product ratio is that it is invariant under row and column multiplication. If we multiply the first and second rows of our fourfold table by $r_1$ and $r_2$ and the first and second columns by $c_1$ and $c_2$, we get the same value for the cross-product ratio that we had in the original table. Because $\alpha$ is invariant under row and column multiplication, this measure of association is not sensitive to the marginal distributions of the variables. A favourable outcome of this is that comparisons can be made between tables which have different marginal distributions. This is in contrast to those $X^2$-based measures of association such as $\phi^2$, $C$, $T$, and $V$. For this reason, and especially because of its clear intuitive interpretation, this ratio is used repeatedly in our study either alone or with other measures of association.

### Yule's $Q$ and $Y$

Two measures of association exist which are attributable to Yule (1912) and which are functions of the cross-product ratio. These are known as Yule's $Q$ and $Y$ and are given by

$$\begin{aligned} Q &= \frac{P_{11}P_{22} - P_{12}P_{21}}{P_{11}P_{22} + P_{12}P_{21}} \\ &= \frac{\alpha - 1}{\alpha + 1} \\ \text{and} \qquad Y &= \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1} \\ \text{so that} \qquad Q &= \frac{2Y}{1 + Y^2} \end{aligned}$$

Except under independence, or strict or weak perfect association when $Q$ or $Y$ are equivalent, $|Y| < |Q|$. The estimates $\hat{Q}$ and $\hat{Y}$ are obtained by replacing $\alpha$ with $\hat{\alpha}$.

Yule's $Q$ and $Y$ have a range of $-1$ to $+1$ with independence between variables resulting in these measures taking a value of zero ($\alpha = 1$). The bounds of $\pm 1$ can again be attained under weak perfect as well as strict perfect association. As previously mentioned, this is not always an attractive feature. Pielou (1969), in specific reference to $Q$, is critical of this measure of association for this reason, claiming that a measure which attains an upper bound under weak perfect association is undesirable, at least in some fields of research such as ecology.

By altering our table for a moment, let us see what this would mean if our cross-tabulation of sex by red meat limitation in the diet were to have had either strict or weak perfect association.

| Sex | Limit Red Meat No | Yes |
|---|---|---|
| Male | 1097 | 0 |
| Female | 0 | 763 |

$\hat{Q} = \hat{Y} = 1$
$\hat{\alpha} = \infty$
$\hat{\phi} = \hat{V}$
$= \hat{T} = 1$
$\hat{C} = .707$

| Sex | Limit Red Meat No | Yes |
|---|---|---|
| Male | 1097 | 0 |
| Female | 1024 | 763 |

$\hat{Q} = \hat{Y} = 1$
$\hat{\alpha} = \infty$
$\hat{\phi} = \hat{V}$
$= \hat{T} = .470$
$\hat{C} = .425$

In the first instance we have strict perfect association with $\hat{\phi} = \hat{V} = \hat{T} = \hat{Q} = \hat{Y} = 1$, $\hat{\alpha} = \infty$, $\hat{C} = .707$ so that all of these measures are attaining their maximum value. That this table depicts strict perfect association is clear. Knowing that a person did not limit the amount of red meat in the diet, we know with certainty that he is male. Similarly, knowing the person limited the amount of red meat in the diet, we know with certainty that she is female.

On the other hand, in our second table, knowing the person responded affirmatively to the question about the restriction of red meat in the diet tells us with certainty that the respondent is female; however, knowing that the person responded negatively to that question does not tell us the sex of that respondent with certainty. In fact, if the respondent answered 'no', the odds that the person is male are 1097 to 1024. We can hardly claim that prior knowledge that the response is 'no' will tell us the person's gender with certainty. Yet while $\hat{\phi}$, $\hat{V}$, $\hat{T}$, and $\hat{C}$ reflect this, $\hat{Q} = \hat{Y} = 1$ and $\hat{\alpha} = \infty$ which indicate perfect association. Clearly the perfect association of this table, however, is not the same as the perfect association of the first.

Two other features which $Q$ and $Y$ have in common with $\alpha$ are that they are invariant under row and column multiplication and under row and column interchange.

Again assuming that the sample size is large, both $Q$ and $Y$ are approximately normally distributed. See, for example, Bishop et al. (1975), Upton (1978), or Liebetrau (1983), for estimates of the means and variances of these measures.

In spite of the similarities of the properties of $Q$ and $Y$ with $\alpha$, the interpre-

tations of these measures are different. To understand the meaning of $Q$ we should note that it is equivalent to Goodman and Kruskal's $\gamma$ (1954) for 2×2 tables. The reader is referred to the later discussion of this measure $\gamma$ for ordinal data, whose interpretation relies on an understanding of concordant (like ordered) and discordant (unlike ordered) pairs of observations from the same population. Davis (1971) gives painstaking details on how to calculate $Q$ and on its intrinsic meaning. In his comparison of $Q$ with $\gamma$, he points out that $Q$ is used when exploring dichotomous variables. While $\gamma$ can also be calculated for 2×2 tables, it should be used with variables with more than two categories which occur 'naturally' as ordinal or are constructed as ordinal from interval or ratio level variables. Fienberg (1977) warns against using Yule's measures when the dichotomous variables are constructed from available continuous bivariate data.

The usefulness of $Y$ as a measure of association appears questionable. Although Bishop, Fienberg and Holland (1975) attempt to interpret $Y$, there does not appear to be any simple meaning of this measure. This fact is acknowledged by others (Reynolds 1977a and Garson 1976, for example), and as Kendall and Stuart (1979) state, "nothing much seems to be gained by the use of $Y$". It is probably for this reason that $Y$ appears infrequently in the literature and also why there is no attempt to discuss its interpretation here as it relates to our data.

## Proportional Reduction in Error Measures

Measures of association exist which have been referred to as proportional reduction in error (or PRE) measures. These measures give the proportional reduction in predictive error which results when one moves from predicting the probability of error in classifying one variable *without* knowledge of the other variable, to the probability of error in classifying the same variable *with* knowledge of the other. With variables

$A$ and $B$, let us denote the two probabilities of error as P(Rule 1) and P(Rule 2) where

P(Rule 1)= probability of error in guessing to which category of one variable (A, say) an individual belongs when the category of the other variable (B, say) to which he belongs is unknown.

P(Rule 2)= probability of error in guessing A when B *is* known.

The PRE measure, $\frac{P(\text{Rule 1})-P(\text{Rule 2})}{P(\text{Rule 1})}$ is the proportional reduction in the amount of error due to knowledge of one variable's category in predicting the category on the other variable.

### Goodman and Kruskal's Lambda, $\lambda$

One such measure, Goodman and Kruskal's $\lambda_A$ gives the following rules for prediction.

**Rule 1:** With no knowledge of $B$ but only of the marginal distribution of $A$, guess that the individual belongs to the $A$ category with the largest marginal probability, denoted $P_{m.} = \max_i\{P_{i.}\}$, $i = 1,\ldots,r$. Therefore, $P_{m.}$ is the probability of correct classification and $1 - P_{m.}$ is the probability of incorrect classification.

**Rule 2:** With knowledge of the $B$ category, guess that the individual belongs to the $A$ category corresponding to the cell with the largest probability in that $j$ column. That is, $P_{mj} = \max_i\{P_{ij}\}$, $i = 1,\ldots,r$ for given $j$ - the maximum cell probability in column $j$ of variable $B$. The probability of error in classification of $A$ given $B$ is $1 - \sum_{j=1}^{c}\max_i\{P_{ij}\} = 1 - \sum_{j=1}^{c} P_{mj}$

Hence,

$$\begin{aligned}\lambda_A &= \frac{P(\text{Rule 1}) - P(\text{Rule 2})}{P(\text{Rule 1})} \\ &= \frac{(1 - P_{m.}) - (1 - \sum_{j=1}^{c} P_{mj})}{1 - P_{m.}} \\ &= \frac{\sum_{j=1}^{c} P_{mj} - P_{m.}}{1 - P_{m.}}\end{aligned}$$

which is estimated by

$$\hat{\lambda}_A = \frac{\sum_{j=1}^{c} f_{mj} - f_{m.}}{n - f_{m.}}$$

where $f_{m.}$ is the maximum marginal total for variable $A$ (rows), $f_{mj}$ is the largest cell frequency in column $j$ of variable $B$ and $n$ is the total sample size.

Just as $\hat{\lambda}_A$ is the estimated proportional reduction in error in predicting $A$ given the predictor or independent variable $B$, the proportional reduction in error in predicting $B$ given the independent variable $A$, is $\hat{\lambda}_B$ where,

$$\hat{\lambda}_B = \frac{\sum_{i=1}^{r} f_{im} - f_{.m}}{n - f_{.m}}.$$

In this way, $\lambda_A$ and $\lambda_B$ are asymmetric measures and might well be employed when one has two-way tables of nominal level variables for which one variable is independent and the other dependent. We may, however, use the following symmetric variation of $\lambda$ in the event that our table is symmetric:

$$\hat{\lambda} = \frac{(\sum_{i=1}^{r} f_{im} - f_{.m}) + (\sum_{j=1}^{c} f_{mj} - f_{m.})}{2n - f_{.m} - f_{m.}}.$$

This is similar in interpretation in that we may think of symmetric $\lambda$ as the proportional reduction in error from knowing the classification of the second variable as opposed to not knowing it. In this instance, however, we do not treat one of the variables as the explanatory and the other as the response variable. Half of the time we estimate the proportional reduction in error in predicting $A$ given $B$ and the other half of the time we estimate the measure for $B$ given $A$.

The following comments apply to $\lambda_A$, $\lambda_B$ and $\lambda$. For simplicity we will refer only to $\lambda_A$. All the $\lambda$ measures vary between 0 and 1. If the variables have no predictive association then the $\lambda$s are zero; the knowledge of classification on the second variable in no way aids in our prediction of the first. In this case the variables are independent. Although the independence of the variables implies that the measure is zero, the converse need not hold. That is, $\lambda_A = 0$ does not necessarily imply that the two variables are independent. (See Upton 1978, or Bishop et al. 1975, who also show this while giving an example of $\lambda$s not being invariant under scale transformation.) If no error is made in guessing $A$ when $B$ is known, then $B$ is a perfect predictor and $\lambda_A = 1$. This will only occur if each column ($B$) has at most one non-zero probability, that is, under strict or implicit perfect association (for columns in this case).

Confidence intervals may be calculated upon estimation of the large sample variance for the $\lambda$ measures. For the formulae for these variances see Goodman and Kruskal (1963), Bishop et al. (1975), Reynolds (1977a), or Liebetrau (1983).

### Goodman and Kruskal's Tau, $\tau$

Another PRE measure due to Goodman and Kruskal is $\tau_A$. The interpretation is the same as before in that it is a proportional reduction in error measurement given by $\frac{P(\text{Rule 1})-P(\text{Rule 2})}{P(\text{Rule 1})}$. The difference is in the rules which are used to classify variable $A$ with and without knowledge of variable $B$.

**Rule 1:** With no knowledge of $B$ but only of the marginal distribution of $A$, classify individuals into categories of $A$ in such a way as to maintain the marginal distribution of $A$. The probability of correct classification in this case is $\sum_{i=1}^{r} P_{i.}^2$ and the probability of incorrect classification is $1 - \sum_{i=1}^{r} P_{i.}^2$.

**Rule 2:** With knowledge of the $B$ category, we again preserve the marginal distribution by classifying an individual into category $i$ of variable $A$ within column $j$ (variable $B$) with probability $\frac{P_{ij}}{P_{.j}}$. The probability of error in classifying $A$ given $B$ is $1 - \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{P_{ij}^2}{P_{.j}}$.

Hence,

$$\begin{aligned}
\tau_A &= \frac{P(\text{Rule 1}) - P(\text{Rule 2})}{P(\text{Rule 1})} \\
&= \frac{(1 - \sum_{i=1}^{r} P_{i.}^2) - (1 - \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{P_{ij}^2}{P_{.j}})}{1 - \sum_{i=1}^{r} P_{i.}^2} \\
&= \frac{\sum_{j=1}^{c} \sum_{i=1}^{r} \frac{P_{ij}^2}{P_{.j}} - \sum_{i=1}^{r} P_{i.}^2}{1 - \sum_{i=1}^{r} P_{i.}^2}
\end{aligned}$$

which is estimated by

$$\hat{\tau}_A = \frac{n \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{f_{ij}^2}{f_{.j}} - \sum_{i=1}^{r} f_{i.}^2}{n^2 - \sum_{i=1}^{r} f_{i.}^2}$$

As with $\lambda$, there is an analogous $\tau_B$ which can be calculated when we wish to see if knowledge of the predictor variable $A$ aids in the classification of $B$. An alternative interpretation of $\tau$ as given by Light and Margolin (based on Gini's work), draws upon analogies with the analysis of variance. See, for example Bishop et al. (1975), Reynolds (1977a), or Liebetrau (1983). This interpretation will not be elaborated on here.

Goodman and Kruskal's $\tau$s vary between 0 and 1. If the variables are independent then $\tau_A = 0$. In the event of perfect prediction (strict or implicit) of $A$ given $B$, $\tau_A = 1$. For formulae for large sample variances of this measure of association, see Goodman and Kruskal (1972).

Both Blalock (1972) and Reynolds (1977a) suggest $\tau$ over $\lambda$ when the marginal distribution of the dependent variable is highly skewed since then $\lambda$ may equal (or

nearly equal) zero – not because the variables are independent but because of the skewness. $\tau$ is not as sensitive to skewed marginals in the dependent variable.

Let us review our table of *health status* versus *smoking habit*. Recall that for this table the $X^2$ (and $G^2$) test for independence was highly significant indicating a relationship between these two variables. Let us examine the strength of this association on the basis of the PRE measure of association.

| | Health Status | | | | |
|---|---|---|---|---|---|
| Smoke | poor | fair | good | excellent | Totals |
| never smoked | 10 | 172 | 696 | 414 | 1292 |
| former smoker | 14 | 125 | 428 | 250 | 817 |
| current smoker | 29 | 251 | 656 | 255 | 1191 |
| Totals | 53 | 548 | 1780 | 919 | 3300 |

$\hat{\lambda}_{hstat} = .04880$
$\hat{\lambda}_{smoke} = .0000$
$\hat{\lambda} = .02778$
$\hat{\tau}_{hstat} = .01066$
$\hat{\tau}_{smoke} = .00616$

The estimated proportional reduction in error in predicting *health status* given the category of *smoking habit* to which the respondents belong is $\hat{\lambda}_{hstat} = .04880$. That is, we reduce the number of errors in classification of *health status* by only 5% by knowing that the person never smoked or is a former or current smoker. Treating *smoking habit* as the dependent variable, we do not reduce the number of errors in classification of this variable at all ($\hat{\lambda}_{smoke} = .0000$) with knowledge of the health status category although, as pointed out, a value of 0 does not necessarily mean that the variables are independent.

If any causal relationship is to be surmised here, we might logically treat *health status* as dependent, working under the assumption that smoking habit influences a person's health status. Others might argue, however, that at least sometimes a person's smoking status might be dictated by his perceived health status with, for example, people giving up smoking when they perceive that their health status is not as good as it should be. If neither variable is treated as the dependent one, we use the symmetric variation of the measure, $\hat{\lambda} = .02778$ and say that by using prior

knowledge about the classification of one variable, we are able to reduce our error in classification by less than 3%.

Using the classification rule of $\tau$, we find that we can reduce our error rate by only 1% ($\hat{\tau}_{hstat} = .01066$) if *health status* is our dependent variable. $\tau$ is more appropriate than $\lambda$ in this case as the dependent variable *health status* is skewed. The smoking variable, however, is not too skewed so there may be no advantage in using $\tau$ over $\lambda$ if *smoking* is treated as the dependent variable. In any case, the values of the measures are so low as to indicate that there is no significant proportional reduction in error regardless of which of the two classification rules is used and regardless of whether we treat this as an asymmetric or symmetric table. That is, there is not a significant reduction in the error rate of classification. Knowledge of smoking habit or health status does not significantly aid in the classification of the other variable.

It is often helpful to have a proportional reduction in error interpretation for a contingency table. Different ones have been examined in our study but given that our health indicator variables are often skewed, we generally use Goodman and Kruskal's $\tau$ measure. This is further illustrated in the loglinear analysis section.

## Measure of Agreement

### Cohen's $\kappa$

One special measure of association is Cohen's $\kappa$ (1960) which measures agreement between two people's categorization or ranking of an item. In our study we use this when we are interested in the degree of agreement between pairs of people such as husbands and wives, as they rank an item. Since each person of the pair rates an item on the same scale, $\kappa$ is used in square $r \times r$ tables only. Cohen's measure is given

by

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{\sum_{i=1}^{r} P_{ii} - \sum_{i=1}^{r} P_{i.} P_{.i}}{1 - \sum_{i=1}^{r} P_{i.} P_{.i}}$$

and estimated by

$$\hat{\kappa} = \frac{n \sum_{i=1}^{r} f_{ii} - \sum_{i=1}^{r} f_{i.} f_{.i}}{n^2 - \sum_{i=1}^{r} f_{i.} f_{.i}}$$

where $P_o = \sum_{i=1}^{r} P_{ii}$ is the proportion of instances in which the pair agrees. This corresponds to those cases appearing on the main diagonal of the table. $P_e = \sum_{i=1}^{r} P_{i.} P_{.i}$ is the proportion of instances of agreement that one would expect to find on the main diagonal by chance under independence. The marginal distributions for the pair must be the same if $\kappa$ is to be able to achieve its maximum (all off-diagonal elements being zero). The division by $1 - P_e$ normalizes $\kappa$ to make it independent of the marginal totals (see Cohen 1960, Reynolds 1977a, or Liebetrau 1983). In the case of perfect agreement we would expect to find all observations on the main diagonal so that $P_o = 1$ and $\kappa = 1$. In the case of independence the amount of agreement is the same as one would expect by chance, so $P_o = P_e$ and $\kappa = 0$. Although independence implies that $\kappa = 0$ the converse need not hold as there may be cases in which association of another kind exists even though agreement does not. See Bishop, Fienberg and Holland (1975) for such an example. If there is absolutely no agreement between the pair, then $P_o = 0$ and $\kappa = -\frac{P_e}{1-P_e}$. Bishop et al. give the estimated asymptotic variance of $\kappa$.

A variation on $\kappa$ is weighted $\kappa$. This considers the case in which one does not simply have agreement or disagreement between pairs of individuals but *degrees* of agreement. For example, those not falling on the main diagonal but belonging to a cell adjoining the main diagonal may show a greater degree of association than those falling far from this diagonal. See Reynolds (1977a) and Liebetrau (1983) for a more complete discussion of weighted $\kappa$.

This is quite a different way of looking at measures of association and one which we use in our study when we wish to ascertain, for instance, the degree of familial agreement. We have already seen measures of association applied to our table of *sex* by *limit red meat*. Now we want to know if husbands and wives responded in the same way to the question of whether or not they consciously limit red meat in their diet for health reasons. It is logical to hypothesize that they would as we would expect them to share many meals and perhaps also share attitudes about diet and nutrition. Here we are only interested in married couples for which both spouses responded to the questi...

Limit Red Meat

| | Husbands | | |
|---|---|---|---|
| Wives | No | Yes | Totals |
| No | 447 | 104 | 551 |
| Yes | 270 | 198 | 468 |
| Totals | 717 | 302 | 1019 |

This gives a value of $\hat{\kappa} = .24076$ which shows some agreement between husband and wife although not as much as one might expect. If then we assume that couples share many meals, we may surmise that in at least some instances, the limitation of red meat in the diet is a *conscious* effort on the part of one member (presumably the one preparing the meal) to have a healthful diet and not a conscious effort on the part of the other spouse. That person may be limiting the amount of red meat in the diet but not for health reasons.

All respondents in this study were asked how satisfied they were with medical care in their own experience on a scale from 1 to 5 where 1 indicated dissatisfaction and 5, satisfaction. It was desirable to see the extent to which married couples responded identically.

Satisfaction With Medical Care

| | Husbands | | | | | |
|---|---|---|---|---|---|---|
| Wives | 1 | 2 | 3 | 4 | 5 | Totals |
| 1 | 2 | 2 | 8 | 6 | 9 | 27 |
| 2 | 5 | 2 | 6 | 4 | 10 | 27 |
| 3 | 5 | 9 | 31 | 46 | 58 | 149 |
| 4 | 9 | 8 | 24 | 90 | 94 | 225 |
| 5 | 17 | 11 | 62 | 112 | 361 | 563 |
| Totals | 38 | 32 | 131 | 258 | 532 | 991 |

This tables gives a value of $\hat{\kappa} = .17020$. Couples do not agree identically most of the time. Note that the marginals are quite skewed here. Most people responded with a level of satisfaction of three or better with the majority of couples tending to be more satisfied than dissatisfied with medical care. We might expect a stronger association if we were to use a weighted $\kappa$. That is, in our calculation we could take into consideration the fact that couples who, for instance, have one partner completely satisfied with health care (5) and the other almost completely satisfied (4) display a greater degree of agreement than the pair with one partner being completely satisfied and the other tending towards dissatisfaction.

## Ordinal Measures of Association

Up to this point the measures of association were primarily for nominal data in two-way contingency tables. We will now concentrate on those tables for which the variables are ordinal. Ordinal variables may arise naturally as in the case of a respondent classifying his health as 'poor', 'fair', 'good', or 'excellent' or may arise when a continuous variable is grouped into discrete categories such as age being grouped as '20-44', '45-64' or '$\geq$65', or as an index of exercise with an underlying continuum being broken down into four categories ranging from 'sedentary' to 'very active'. Clearly there is a loss of information when we categorize variables yet such categorization is necessary if we wish to analyze the data using measures of association in contingency tables or loglinear analysis. Many of the variables in this study are

ordinal either naturally or due to grouping, hence the measures of association in this section are important. This is not to say that we cannot use previously discussed measures. Indeed these may be used even when we have ordinal variables but since they ignore ordinality they do not take advantage of all the available information.

### Measures Based on Concordance and Discordance

There are several measures of association which are based on the difference in concordant and discordant pairs in a contingency table in which both variables are ordered. Rather than considering individuals, we must now think in terms of pairs of individuals drawn at random. The pair is called concordant if one of the individuals ranks higher than the other individual on both variables $A$ and $B$. The pair is discordant if they rank in opposite directions, that is, if an individual in the pair ranks higher than the other individual on one variable but lower on the other variable. The remaining possibility is that the pair is tied. This can happen in three ways; the pair may be tied on variable $A$ but not on variable $B$, on $B$ but not on $A$, or on both variables. The following notation is used:

$$
\begin{array}{lcl}
C & = & \text{the number of concordant pairs} \\
D & = & \text{the number of discordant pairs} \\
T_A & = & \text{the number of pairs tied on } A \text{ but not on } B \\
T_B & = & \text{the number of pairs tied on } B \text{ but not on } A \\
T_{AB} & = & \text{the number of pairs tied on } A \text{ and } B
\end{array}
$$

The total number of possible pairs is $\binom{n}{2}$ where $n$ is the number of individuals in the table.

Formulae for calculating these pairs may be found in numerous texts (Kendall and Stuart 1979, Hays 1981, Agresti 1984, or Freeman 1987).

**Kendall's $\tau$s**

Kendall considered the difference between the probability of occurrence of concordant pairs ($P_C$) and the probability of discordant pairs ($P_D$) with his measure of association, $\tau = P_C - P_D$. This measure was constructed under the assumption that the variables in question were continuous and could be completely ranked with no ties occurring in the pairs.

An estimate of $\tau$ is given by $\hat{\tau}_a = (C - D)/\binom{n}{2}$ where $\binom{n}{2}$ includes *all* possible pairs, whether or not tied. Thus this measure can be interpreted as the difference in the concordant and discordant pairs over all possible pairs of individuals drawn at random. This attains the extreme value of 1 when all the pairs are concordant and $-1$ when all the pairs are discordant. When $A$ and $B$ are independent, the chance of having concordant pairs is the same as that of having discordant pairs so the measure is zero. The converse of this need not hold. This measure of association is relatively easy to understand but should be used with caution in contingency table analysis since the assumption of continuous variables is violated and ties *do* exist.

Two alternative $\tau$ measures exist. Their estimates are denoted $\hat{\tau}_b$ and $\hat{\tau}_c$ where

$$\hat{\tau}_b = \frac{C - D}{\sqrt{(C + D + T_A)(C + D + T_B)}}$$

and

$$\hat{\tau}_c = \frac{C - D}{n^2(m-1)/2m} \quad \text{where} \quad m = \min\{r,c\}$$

See Liebetrau (1983) for estimated large sample variances of $\hat{\tau}_a$, $\hat{\tau}_b$ and $\hat{\tau}_c$.

$\tau_b$ compensates somewhat for the fact that ties exist in our two-way tables. The extreme values of $\pm 1$ are attained by $\tau_b$ when all the pairs of observations are concordant ($\tau_b = 1$) or discordant ($\tau_b = -1$). This can only happen when the contingency table is square so that $\tau_b$ cannot attain its maximum if $r \neq c$. As before,

if the variables are independent this measure is zero but the converse is not necessarily true.

Stuart (1953) modified $\tau$ with $\tau_c$ to allow a non-square contingency table to attain its extreme values. Several authors have pointed out the difficulty in interpreting this measure, however.

Generally speaking these measures of association are dependent upon the number of categories present in the table. As the number of categories increases, the number of ties should decrease and therefore the closer the estimated measure should approach the true difference in the proportion of concordant and discor 'ant pairs. In the presence of many ties, these measures tend to understate the degree of association.

### Goodman and Kruskal's Gamma, $\gamma$

Goodman and Kruskal's $\gamma$ (1954) is a measure of association which is also based on concordant and discordant pairs. Given by

$$\gamma = \frac{P_C - P_D}{P_C + P_D} \quad \text{and estimated by} \quad \hat{\gamma} = \frac{C - D}{C + D}$$

it is the difference in the probabilities of concordant and discordant pairs conditional on there being no ties at all. If the two variables are independent, $\gamma = 0$ but the converse need not hold. It not only attains its extreme values of $\pm 1$ under strict perfect positive or negative correlation but also under asymmetric perfect or weak perfect correlation. See Reynolds (1977a) for illustrations of these different types of correlation. For the formula for the estimated large sample variance of $\hat{\gamma}$, see Liebetrau (1983).

If there are a lot of ties then $\gamma$ tends to overstate the true measure of association; the more ties, the greater the degree to which this is true. This might be the case

especially as the number of categories in a table decreases as, all else being equal, the proportion of ties will increase as the number of categories decreases. Although this is a problem with $\gamma$, it is an appealing measure due to its simple interpretation. As already mentioned, it can be interpreted in terms of the difference in proportions of concordant and discordant pairs. As explained by Mueller et al. (1977) in some detail and by Costner (1965), it can also be interpreted as a proportional reduction in error measure.

As a PRE measure, we are interested in the prediction of order for pairs. All ties, of course, are still ignored. The rules follow:

**Rule 1:** With no knowledge of the order for a pair on independent variable $B$, we guess the order for that pair on dependent variable $A$. When we draw a pair, we guess that the first unit of the pair is the higher on $A$. The probability of error in the prediction then, is $\frac{1}{2}$ and the estimated number of errors is $\frac{1}{2}(C + D)$.

**Rule 2:** With knowledge of the order for the pair on variable $B$, guess that the order for that pair on variable $A$ is the same as for $B$ if the number of concordant pairs is greater than the number of discordant pairs; guess that the order for that pair on $A$ is the opposite of the order on $B$ if the number of concordant pairs is less than the number of discordant pairs. In other words, for each pair drawn, guess concordance if $C > D$ and discordance if $C < D$. The estimated number of errors is $\min(C, D)$. Hence with

$$\gamma = \frac{P(\text{Rule 1}) - P(\text{Rule 2})}{P(\text{Rule 1})}$$

we have

$$\hat{\gamma} = \frac{\frac{1}{2}(C + D) - \min(C, D)}{\frac{1}{2}(C + D)}$$

$\gamma$ is the proportional reduction in error in predicting the order of pairs when rule 2 is used in lieu of rule 1. So in spite of its drawbacks we see that $\gamma$ has two intuitively pleasing interpretations. Some authors suggest that if $\gamma$ is to be used, other measures of association should be reported with it.

Because of its interpretation we frequently use this measure in our health study when we have tables which have ordinal variables. With such tables we use other measures alongside $\gamma$. Examples are given in a later section.

### Somers' $d$ and Wilson's $e$

Yet another measure based on concordant and discordant pairs is Somers' (1962) asymmetric measure, $d$. Somers' $d$ is the difference in the probabilities of concordant and discordant pairs assuming there are no ties whatsoever on the independent variable. Although Costner (1965) states that Somers' asymmetric measure has no proportional reduction in error interpretation, Reynolds (1977a) (with minor modification to rule 2) tries to give such an explanation to the absolute value of this measure. Denoted $d_A$ when $A$ is the dependent variable and $B$ is the independent variable, the estimate is given by

$$\hat{d}_A = \frac{C - D}{C + D + T_A}$$

Similarly, $\hat{d}_B = \frac{C-D}{C+D+T_B}$ is the estimate for $d_B$ where $B$ and $A$ are the dependent and independent variables, respectively.

When $A$ and $B$ are independent, $d_A$ (or $d_B$) is zero. The extreme values of $\pm 1$ can be reached in non-square as well as square tables although as Reynolds (1977a) points out, when the table is not square the maximum is attainable only when the variable with the fewer categories is the independent variable. Goodman and Kruskal (1972) give the asymptotic variance of Somers' asymmetric measure.

Liebetrau (1983) and Garson (1976) also give a symmetric measure of $d$ which is estimated by

$$\hat{d}_{sym} = \frac{C - D}{C + D + \frac{1}{2}(T_A + T_B)}$$

Yet another similar measure is Wilson's $e$ (1974). A symmetric measure, this looks at the difference in the probabilities of concordant and discordant pairs given there are no pairs tied on *both* variables. The estimate is given by

$$\hat{e} = \frac{C - D}{C + D + T_A + T_B}$$

If $A$ and $B$ are independent, $e = 0$. The measure can only attain extreme values of $\pm 1$ in square tables where there are no ties on $A$ alone and none on $B$ alone.

Let us review again our table of *smoking* by *health status*. Although we have already discussed this cross-tabulation in terms of several measures of association, in the previous discussion we treated the variables as nominal. This is acceptable but since we ignored the fact that both variables are ordinal, we did not avail of all the information on hand.

| | Health Status | | | | |
|---|---|---|---|---|---|
| Smoke | poor | fair | good | excellent | Totals |
| never smoked | 10 | 172 | 696 | 414 | 1292 |
| former smoker | 14 | 125 | 428 | 250 | 817 |
| current smoker | 29 | 251 | 656 | 255 | 1191 |
| Totals | 53 | 548 | 1780 | 919 | 3300 |

$\hat{\tau}_b = -.1154$
$\hat{\tau}_c = -.1088$
$\hat{\gamma} = -.1827$
$\hat{d}_{sym} = -.1153$
$\hat{d}_{smoke} = -.1202$
$\hat{d}_{hstat} = -.1107$

$\hat{\tau}_b = -.1154$ implies that there is a small degree of negative association. That is, we have more discordant than concordant pairs; if an individual in the pair ranks higher than the other person on the smoking variable, then that individual is more likely to rank lower on the health status variable. This is the trend we would expect to see. Although we report it here, in our study we prefer other ordinal measures since the interpretation of this measure is difficult.

In terms of discordant and concordant pairs, $\hat{\gamma} = -.1827$ also indicates a negative association between the variables. Although the degree of association is somewhat larger with this measure, it must be remembered that ties are present yet the measure does not allow for ties, therefore any value of $\gamma$ is likely to overstate the true degree of association. Using a PRE measure interpretation we say that the absolute value of $\hat{\gamma}$ implies that with knowledge of the number of discordant and concordant pairs we reduce the percent of errors of classification of the pair by over 18% from what we would have without this knowledge. With these two useful interpretations, we favour this measure above many of the others and use it quite extensively in our health study when we have contingency tables which have ordinal variables. Due to its exaggerated value when there are ties, however, we do not report this as the sole measure.

A case may be made for the our treating the smoking variable as dependent upon health status although if any causal relation is assumed, the reverse is the more acceptable. Regardless of the order of causality – if indeed it is to even be thought of as asymmetric – the value for Somers' measure of association takes a value between $-.12$ and $-.11$ so that we again say that there is weak negative correlation between the variables. Similar to $\gamma$, if we give this a PRE interpretation we say that by moving from not knowing to knowing the order of a pair on the independent variable, *smoking* say, we reduce the percent of errors in predicting the correct order on the dependent variable, *health status*, by 11% by predicting the order based on the number of concordant and discordant pairs. Likewise, a similar interpretation may be given if we treat *smoking* as the dependent variable.

A number of measures of association have been mentioned in this chapter. There are others such as Mantel-Haenszel, tetrachoric correlation, McNemar's test, uncertainty coefficient and Spearman's rank correlation to name a few. These shall not be

discussed. Those dealt with were chosen because of their usefulness as measures in this study or because of their inherent interest.

## 3.2 Design Effects

### 3.2.1 The Design Effect in a 2×2 Health Table

The analyses that are used in this study assume that the data were collected by simple random sampling and the multinomial sampling model. As with most surveys, however, the sampling method – single-stage cluster design – was somewhat more complex than this. When analytical techniques which assume a simple sampling design are used to study data collected under more complex schemes, a clear violation of an assumption has taken place. Before assuming that one's results are acceptable then, the researcher should examine how serious this violation is. That is, he should look at the design effects, or *deffs*.

The *deff* is the ratio of the variance estimates under the sampling design to those estimates under simple random sampling. Clearly then, if there is no design effect, the ratio will be unity. The greater the design effect, the further this value will be from 1. If *deff*>1, then by using formulae for simple random sampling instead of for clustering we are underestimating the variance for the variable. Likewise, if *deff* <1 we are overestimating it.

Since our data were collected using a single-stage cluster design, we must acknowledge possible dependency within sampling units or households and must consider the design effects. In our analysis we have been dealing with categorical data and have been examining cross-tabulations of variables in some detail, so we now look at *deffs* for proportions appearing in cells of contingency tables of discrete variables. We explore the effect that dependency among sampling units has on the familiar $\chi^2$ test statistic, $X^2$, as used to test for the independence between variables. We will examine the design effects in this context and explore possible correction factors for

the statistic. We should bear in mind that in light of our previous discussion, $X^2$ may not be the best statistic to use, corrected or not. Earlier in this chapter we explored the use of other statistics as measures of association.

The following table of interest to medical researchers will suffice as an illustration. We will first briefly consider a 2×2 table of *health status* (HS) by *health practices* (HP) and later will look with more detail at the 2×7 table from which this was obtained. The health practices are those used in the Alameda County Survey (Belloc and Breslow 1972, Belloc 1973, Breslow and Enstrom 1980) namely, eating breakfast, number of hours sleep, number of alcoholic drinks, smoking, weight, and exercise.

Consider the table below. Note that both variables have been dichotomized. HS takes a value of 0 (poor or fair) or 1 (good or excellent) while HP assumes a value of 0 (0-1 health practices) or 1 (2-6 health practices). This particular dichotomization of the original table resulted from epidemiological considerations.

| Health Status (HS) | Health Practices (HP) | | |
|---|---|---|---|
| | 0 - 1 | 2 - 6 | Totals |
| 0 | 56 | 533 | 589 |
| 1 | 162 | 2497 | 2659 |
| Totals | 218 | 3030 | 3248 |

Let us use the following standard notation for this example where we concentrate on those respondents belonging to the first cell.

$a_i =$ number of respondents in the $i^{\text{th}}$ cluster with poor or fair health status and 0-1 health practices

$m_i =$ size of the $i^{\text{th}}$ cluster

So the proportion of those in the sample with poor or fair health status and with 0-1 health practices is

$$p = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i}, \quad n = \text{number of clusters}$$

The variance estimate of $p$ under binomial theory is

$$\hat{V}_{binomial}(p) = \hat{V}_b(p) = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i}\left(1 - \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i}\right) / \sum_{i=1}^n m_i = p(1-p)/n$$

And under single-stage cluster sampling, the variance is

$$\hat{V}_{cluster}(p) = \hat{V}_c(p) = \frac{1}{n\bar{m}^2}\frac{\sum_{i=1}^n a_i^2 - 2p\sum_{i=1}^n a_i m_i + p^2\sum_{i=1}^n m_i^2}{n-1}$$

$$\bar{m} = \frac{\sum_{i=1}^n m_i}{n}, \quad p = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i}$$

In our example we have the following values:

$$\begin{aligned}
n &= 1648 \\
\textstyle\sum_{i=1}^n a_i &= 56 \\
\textstyle\sum_{i=1}^n m_i &= 3248 \\
\textstyle\sum_{i=1}^n a_i m_i &= 132 \\
\textstyle\sum_{i=1}^n a_i^2 &= 58 \\
\textstyle\sum_{i=1}^n m_i^2 &= 7566
\end{aligned}$$

Note that $n$ will change depending on the table.

The variance estimates are

$$\begin{aligned}
\hat{V}_b(p) &= 5.2167839\times10^{-6} \\
\hat{V}_c(p) &= 5.2828267\times10^{-6}
\end{aligned}$$

to give an estimated design effect of

$$deff = \frac{\hat{V}_c(p)}{\hat{V}_b(p)} = 1.0126597$$

So by using the formula for simple random sampling we are underestimating the variance for the proportion for this cell. Given the proximity of this value to unity, if the sizes of the *deffs* for the other cells in the contingency table are in keeping with this *deff*, it would seem reasonable to say that in this example our $\chi^2$ test statistic, $X^2$, will not be unduly affected by our assumption of a multinomial, instead of a

cluster sample. Fellegi (1978) mentions that *deff* is dependent upon several factors and that in "well-designed surveys it ranges typically between 1 and 3 ... [with] the most common values appear[ing] to be between 1.4 and 2". So then with this in mind, our observed *deff* of 1.01 is certainly an acceptable value.

In general we note that the dependency among household members may be such that a correlation exists. The intraclass correlation coefficient, $\rho$, is the correlation between all the possible pairs of elements within clusters. The formula for *deff* can be written in terms of this coefficient. That is, we express the estimated design effect as follows:

$$deff = \frac{\hat{V}_c(p)}{\hat{V}_b(p)} = \rho(\bar{m} - 1) + 1$$

so that $\rho$ is

$$\rho = \frac{\hat{V}_c(p) - \hat{V}_b(p)}{\hat{V}_b(p)(\bar{m} - 1)}$$

When we have independence among household members, we may ignore the fact that we have clusters. If there is no correlation within the clusters, then $\rho = 0$ and $deff = \rho(\bar{m} - 1) + 1 = 1$, clearly regardless of the average cluster size. From the estimates for one cell in our illustration, we calculate the estimate, $\rho = 0.0130394$. As Cochran (1977) notes, since $\rho > 0$, the estimated variances reflect that the use of cluster sampling here is less precise than that of simple random sampling although in this case it is marginally less. In another table we might expect to see a larger $\rho$; we would anticipate that households exhibit varying strengths of intraclass correlation coefficients depending upon the variables under examination. Sudman (1976) has a nice discussion of the interpretation of the intraclass correlation coefficient under cluster sampling. He includes in his discussion, a table of measures of $\rho$ from the National Health Survey (Source: U.S. National Center for Health Statistics) for average cluster sizes of 6, 9 and 18 where 'cluster size' refers to the number of households

in a cluster and where all members of a household were surveyed. In summarizing this table, Sudman states that "in general, values of $\rho$ for health statistics are small, averaging around .05 or lower". This agrees with the estimated value of $\rho$ for our table.

When we express *deff* in terms of $\rho$ it is apparent that, all else being equal, the closer the average cluster size $\bar{m}$ is to one, the closer is *deff* to unity with it equalling unity if $\bar{m}$ is one. This is obvious since single-stage cluster sampling reduces to simple random sampling if the average cluster size is one. As $\bar{m}$ increases, even with a small correlation coefficient the design effect's departure from unity increases. In our example, $\bar{m} = 1.97$. Cochran notes that we would expect the variance calculated between members in the same household to grow as the size of the cluster grows. An average cluster size of approximately two, such as we have in our health survey, is not large. Even so, we must consider it together with the intraclass correlation coefficient when determining how serious we regard any deviation from one.

### 3.2.2 Using Design Effects to Correct for $X^2$ in a 2×7 Health Table

Ideally we desire design effects of unity. Provided that the *deffs* are close to this we can proceed with our analysis without any grave misgivings, using formulae for simple random sampling instead of for clustering. The question remains as to what we can do in the event that the *deffs* are not deemed negligible. Then the effect on $\chi^2$ test statistics of assuming simple random instead of cluster sampling should not be ignored. The least we would want to do in such a case is apply a correction factor to the test statistic. We continue now to show how to use calculated *deffs* as corrections to $X^2$.

Much has been written about how $X^2$ may be corrected in tests of independence in $r \times c$ tables. In the literature the Wald statistic, which is distributed asymptotically as $\chi^2_{(r-1)(c-1)}$, is suggested as an appropriate statistic since it may be used even under complex survey designs. See Rao and Scott (1981, 1984), for example, for a discussion of the Wald statistic. The variance-covariance matrix of cell estimates which is required for the calculation of this statistic is not always readily available although it can be calculated when the primary data is available. Fellegi (1978) comments on the necessity of making strong simplifying assumptions in order to estimate covariance matrices in complex surveys.

There has been, in the literature, some discussion on particular covariance structures. Cohen (1976) examines a model of clustering which allows for positive association only and which has clusters of units each of size two. He provides the covariance matrix for this particular model. Altham (1976) extends Cohen's results. Whereas Cohen considered family clusters of size two, Altham examines those of size $k$ and gives the resulting covariance matrix for this somewhat more complex model. The clusters, however, are still of a constant size. Brier (1980) takes this one step further by looking at clusters of unequal sizes as well as those of equal sizes. He does so by assuming a Dirichlet-multinomial distribution as a model. The covariance matrix resulting from this assumption is discussed in his 1980 paper. Also see Fingleton's (1984) synopsis of Brier's paper. Thomas and Rao (1987) discuss four adjustments to $X^2$ in tests of goodness-of-fit in cluster sampling and comment upon their comparative value. They look at a modified Wald statistic, Fay's jackknifed $X^2$, and two corrections proposed by Rao and Scott. With the exception of one of Rao and Scott's statistics which relies only upon knowledge of the estimated cell variances, the aforementioned statistics discussed in this paper require the covariance matrix.

Rao and Scott (1981) show that for tests of independence in an $r \times c$ table,

a correction to $X^2$ can be made from knowing only the cell proportions and the estimated *deffs* of these cell proportions and marginals. As discussed in their 1984 paper, in three-way tables correction factors can be expressed in terms of the cells' proportions, their estimated *deffs* and the estimated *deffs* of the one and two-way marginals, depending on the hypothesis under study. It can be expressed this way, for example, when the hypothesis is of complete independence. For other hypotheses, however, a more complicated procedure involving estimation of the full covariance matrix, which is often not available, may be required. For the purpose of this report, we have investigated the $r \times c$ table under the hypothesis of independence. We have found such a minimal clustering effect with cluster sizes which are very small that we will ignore the effect of the survey design. This has been the usual practice and will probably continue to be until computer programs which calculate variance and covariance estimates become readily obtainable. This appears to be a safe practice when examining health variables.

Rather than examine statistics which depend upon knowledge of the covariance matrix, we concentrate on two corrections to $X^2$ which have been proposed by Fellegi (1978, 1980) and by Rao and Scott (1981, 1984) for tests such as the test of independence in a two-way table. That Fellegi's and Rao and Scott's tests do not require knowing the covariance structure in this instance, makes them more readily calculable than some of the other proposed statistics. It has been pointed out by Holt, Scott and Ewings (1980), however, that these tests, as well as those put forth by Cohen, Altham, Brier and others, are conservative in tests for independence and perform less well than when used in tests of goodness-of-fit. Still, their *comparative* facility of calculation makes them worthy of consideration. They were calculated by means of a FORTRAN program written by this author. This program calculates the *deffs* for a 2×7 table from our single-stage cluster sample of 1648 clusters of approx-

imate average size of two. Note that the tedious nature of the programming required to produce the output will probably prevent most researchers intending to use log-linear analysis, for instance, from computing the design effects. This should change as programs that are easily adaptable to perform the calculations required under the design at hand become readily available. Programs that calculate variance estimates for data collected from complex survey designs do exist but were unavailable to the author. These include SUDAAN, distributed by Research Triangle Institute in North Carolina, and SUPER CARP from Iowa State University.

Before proceeding, let us review the notation which we will require for the discussion of these corrections:

Let $Y_{ijkl} = 1$ if the $l^{\text{th}}$ observation in the $i^{\text{th}}$ cluster belongs to the $jk^{\text{th}}$ category

$= 0$ otherwise

where, $i = 1, 2, ..., n$ $n$ =number of clusters

$j = 1, 2, ..., r$ $r$ =number of rows

$k = 1, 2, ..., c$ $c$ =number of columns

$l = 1, 2, ..., m_i$ $m_i$ =number of respondents in the $i$th cluster

$Y_i = \sum_l Y_{ijkl}$ = number in the $jk^{\text{th}}$ category from $i^{\text{th}}$ cluster

$Y_i' = \sum_k \sum_l Y_{ijkl}$ = number in the $j^{\text{th}}$ row of the $i^{\text{th}}$ cluster

$Y_i'' = \sum_j \sum_l Y_{ijkl}$ = number in the $k^{\text{th}}$ column of the $i^{\text{th}}$ cluster

$n_{jk} = \sum_i \sum_l Y_{ijkl} = \sum_i Y_i$ = number in the $jk^{\text{th}}$ category

$\bar{m} = \frac{\sum_i m_i}{n} = \frac{N_T}{n}$ = average cluster size,
$N_T$ = total number of respondents

$$p_{ijk.} = \frac{\sum_l Y_{ijkl}}{m_i} = \frac{Y_i}{m_i} = \text{proportion in the } jk^{\text{th}} \text{ category of the } i^{\text{th}} \text{ cluster}$$

$$p_{.jk.} = \frac{\sum_i \sum_l Y_{ijkl}}{\sum_i m_i} = \frac{n_{jk}}{N_T} = \text{proportion in the } jk^{\text{th}} \text{ category over all clusters}$$

$$p_{.j..} = \frac{\sum_k \sum_i \sum_l Y_{ijkl}}{\sum_i m_i} = \frac{\sum_k \sum_i Y_i}{N_T} = \frac{\sum_k n_{jk}}{N_T} = \frac{n_{j.}}{N_T}$$
$$= \text{proportion of the } j^{\text{th}} \text{ row marginal}$$

$$p_{..k.} = \frac{\sum_j \sum_i \sum_l Y_{ijkl}}{\sum_i m_i} = \frac{\sum_j \sum_i Y_i}{N_T} = \frac{\sum_j n_{jk}}{N_T} = \frac{n_{.k}}{N_T}$$
$$= \text{proportion of the } k^{\text{th}} \text{ column marginal}$$

Rao and Scott describe how $X^2$ may be corrected in tests of independence in $r \times c$ tables. Their correction factor, $\bar{d}$, relies upon the knowledge of the design effect, $d_{jk}$, for each cell in the contingency table and upon the design effects, $d_j(r)$ and $d_k(c)$, of the row and column marginals of that table. Unlike the Wald statistic, it does not require knowledge of the full covariance matrix of cell estimates. The calculated $d_{jk}$s are the ratios of variance estimates of the cell proportions under cluster sampling to the variance estimates under multinomial sampling. These estimated cell design effects are given by

$$d_{jk} = \frac{\hat{V}_c(p_{.jk.})}{\hat{V}_l(p_{.jk.})} = \frac{N_T}{n\bar{m}^2(n-1)} \frac{\sum_{i=1}^n (Y_i - p_{.jk.} m_i)^2}{p_{.jk.} q_{.jk.}} = \frac{1}{\bar{m}(n-1)} \frac{\sum_{i=1}^n (Y_i - p_{.jk.} m_i)^2}{p_{.jk.} q_{.jk.}}$$

since

$$\hat{V}(p_{.jk.})_{binomial} = \hat{V}_b(p_{.jk.}) = \frac{p_{.jk.} q_{.jk.}}{N_T}, \quad q_{.jk.} = 1 - p_{.jk.}$$

and

$$\hat{V}(p_{.jk.})_{cluster} = \hat{V}_c(p_{.jk.}) = \frac{\sum_{i=1}^n (Y_i - p_{.jk.} m_i)^2}{n\bar{m}^2(n-1)}.$$

The *deffs* for row and column marginals, $d_j(r)$ and $d_k(c)$ respectively, are

$$d_j(r) = \frac{1}{\bar{m}(n-1)} \frac{\sum_{i=1}^n (Y_i' - p_{.j..} m_i)^2}{p_{.j..} q_{.j..}}$$

and

$$d_k(c) = \frac{1}{\bar{m}(n-1)} \frac{\sum_{i=1}^{n}(Y_i'' - p_{..k.}m_i)^2}{p_{..k.}q_{..k.}}$$

Finally we calculate $\bar{d}$ as

$$\bar{d} = \frac{1}{(r-1)(c-1)} \sum_{j=1}^{r}\sum_{k=1}^{c}(1-p_{.j.}p_{..k.})d_{jk} - \sum_{j=1}^{r}(1-p_{.j..})d_j(r) - \sum_{k=1}^{c}(1-p_{..k.})d_k(c)$$

Fellegi proposed a somewhat simpler correction, $\bar{\bar{d}}$, which requires the matrix of variance estimates for the cell proportions but does not rely upon the knowledge of the row and column marginals. This factor, being the average of the *deffs* of proportion estimates in the $jk^{\text{th}}$ category over all clusters, is given by

$$\bar{\bar{d}} = \frac{1}{rc}\sum_{j=1}^{r}\sum_{k=1}^{c} d_{jk}$$

We will apply these formulae to a 2×7 health table. The cell and marginal *deffs* as well as the correction factors were calculated using the aforementioned FORTRAN program. In our previous example we cross-tabulated *health status* (HS) and *health practices* (HP) after both variables were dichotomized. Now we consider HS using the same dichotomy as before, but treat HP as the number of health practices (0 to 6, inclusive) existing in the original variable prior to any recoding. That table is given below for both the frequencies of occurrence and for the proportions.

**Frequencies of HS by HP**

| | HP | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| HS | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Totals |
| 0 | 9 | 47 | 139 | 199 | 123 | 63 | 9 | 589 |
| 1 | 22 | 140 | 451 | 828 | 751 | 386 | 81 | 2659 |
| Totals | 31 | 187 | 590 | 1027 | 874 | 449 | 90 | 3248 |

**Proportions of HS by HP**

| | HP | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| HS | 0 | 1 | 2 | 3 | 4 | 5 | 6 | |
| 0 | .0027709 | .0144704 | .0427956 | .0612685 | .0378695 | .0193966 | .0027709 | .1813424 |
| 1 | .0067734 | .0431034 | .1388547 | .2549261 | .2312192 | .1188424 | .0249384 | .8186576 |
| | .0095443 | .0575738 | .1816503 | .3161946 | .2690887 | .1382390 | .0277093 | 1 |

Recall that under simple random sampling the $\chi^2$ test statistic, $X^2$, for an $r \times c$ table is given by

$$X^2 = N_T \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - \hat{P}_{i.}\hat{P}_{.j})^2}{\hat{P}_{i.}\hat{P}_{.j}}$$

where $p_{ij}$ is the observed proportion in the $ij^{\text{th}}$ cell, and $\hat{P}_{i.}$ and $\hat{P}_{.j}$ are the estimated expected proportions for the row and column marginals, respectively and $X^2 \sim \chi^2_{(r-1)(c-1)}$. In our example, $X^2$= 40.227980 with df=6 so we reject our hypothesis of independence.

As given earlier, under simple random sampling, to calculate the estimated variance of a proportion, $\hat{V}_b(p)$ in any given cell we compute values by the binomial formula, $p(1-p)/n$ where $p$ is the proportion in the cell and $n$ is the total number of individuals in the sample. Applying this formula to our data yields the next table:

**Variances of proportions under SRS (Multinomial/Binomial) Sampling**

| | HP | | | | | | |
|---|---|---|---|---|---|---|---|
| HS | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | .0000009 | .0000044 | .0000126 | .0000177 | .0000112 | .0000059 | .0000009 |
| 1 | .0000021 | .0000127 | .0000368 | .0000585 | .0000547 | .0000322 | .0000075 |

The formula for calculating the estimate of the variance of a proportion under cluster sampling, $\hat{V}_c(p)$, was also given earlier and the cross-tabulation of these values follow:

**Variances of proportions under Cluster Sampling**

| | HP | | | | | | |
|---|---|---|---|---|---|---|---|
| HS | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | .0000008 | .0000045 | .0000125 | .0000178 | .0000110 | .0000062 | .0000008 |
| 1 | .0000020 | .0000125 | .0000390 | .0000594 | .0000603 | .0000360 | .0000077 |

The cell *deffs* are given below:

**Design Effects, $d_{jk}$: The ratio of the variance of the proportions under Cluster Sampling to the variance under Multinomial Sampling for the individual cells**

| | HP | | | | | | |
|---|---|---|---|---|---|---|---|
| HS | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | .9968768 | 1.0233425 | .9908524 | 1.0070701 | .9804897 | 1.0571659 | .9987280 |
| 1 | .9885748 | .9830058 | 1.0596248 | 1.0157695 | 1.1100599 | 1.1174185 | 1.0297006 |

The correction factors due to Rao and Scott ($\overline{d}$) and Fellegi ($\overline{\overline{d}}$) are

$$\overline{d} = 0.995075 \qquad \text{and} \qquad \overline{\overline{d}} = 1.025620$$

so that our corrected $X^2$s are

$$X^2_{\overline{d}} = 40.427083 \qquad \text{and} \qquad X^2_{\overline{\overline{d}}} = 39.223085$$

Clearly these statistics are so close in value to the uncorrected $X^2$ that they do not change our conclusion that we reject our hypothesis of independence. At a glance, we have the following:

| Variables | df | $\overline{d}$ | $\overline{\overline{d}}$ | $X^2$ | $X^2_{\overline{d}}$ | $X^2_{\overline{\overline{d}}}$ |
|---|---|---|---|---|---|---|
| HP×HS | 6 | 0.995075 | 1.025620 | 40.227980 | 40.427083 | 39.223085 |

In summary, our average cluster is only approximately of size two and we have a reasonably large sample size. All else being equal, the seriousness of the *deffs* may be greater for larger cluster sizes or particular covariance matrices (Rao and Scott 1981, Thomas and Rao 1987). The cell *deffs* are small as are the marginal *deffs*. With respect to the marginal *deffs*, Holt, Scott and Ewings (1980) warn against using $X^2$ test statistics without some correction factor if, in a two-way table, both variables have marginals with high *deffs*. In our illustration the design effects are such that neither our cell nor marginal *deffs* should cause us undue alarm as in no instance were they as large as 1.2. Given this combination of factors, we conclude that it is not necessary for us to apply correction factors to $X^2$ in this instance.

It might be noted here that when future studies of a similar nature are carried out by this research team, it would be worthwhile to calculate *deffs* for those variables in contingency tables which will be examined. Design effects allow us to judge whether or not it is reasonable to proceed with analyses which assume simple random sampling. So too, we may use design effects as inflation factors to the sample size if subsequent studies are going to again employ cluster sampling of households. That is, we calculate the sample size under the assumption of simple random sampling and then multiply by this factor (Cochran 1977).

## 3.3 Logistic Regression for Health Status and Two Health Practices

Although analysis by cross-tabulation using measures of association is interesting and has a place in trying to get a profile of our sample, there are other analytical tools available which allow us to explore our data further. The research questions posed, together with the nature of much of the data, led to logistic regression and loglinear analysis being among the tools used in this study.

In several instances we wished to examine how variables which are thought to be health indicators, are related to a dichotomous self-assessed health status variable which takes a value of 0 if health is poor (fair/poor) or 1 if it is good (good/excellent). These variables are behaviors or practices whose presence or absence, or degree thereof, have been regarded as indicators of overall health status in previous studies. See, for example, the Alameda County Study as mentioned earlier.

### 3.3.1 Sleep and Health Status

There are many different relationships between our health variables whose investigation is worthwhile. One hypothesis was that a moderate amount of sleep is associated with good self-assessed health status. More specifically, if sleep is associated with health status, what is the optimum number of hours of sleep?

Excluding obvious outliers, such as an average of 1 or 20 hours of sleep per night, our independent variable sleep took values between 3 and 10 and our binary response variable took values of 0 or 1, for poor or good self-assessed health status, respectively. Using logistic regression to explore the relationship between these variables, we fit a

logistic regression model of the standard form

$$P(success) = \frac{e^u}{1+e^u}$$

where $P(success)$ is the estimated probability that the respondent will have good health status (success, $Y_{ghstat} = 1$) rather than poor (failure, $Y_{ghstat} = 0$). $u$ is a linear function of the independent variable, sleep. That is, $u = \beta_0 + \beta_1 X_{sleep}$. Expressed in terms of the logit, or log of the odds, $u = \log\left(\frac{P(success)}{1-P(success)}\right)$.

We begin by fitting a simple linear model and plan to move to a more complex model if it is needed. Using the statistical computer package BMDP, a stepwise logistic program LR, was run to fit a linear model in the variable sleep as described above. In the BMDP program, a model was specified with the interval variable $X_{sleep}$ as the independent variable and the grouped self-assessed health status variable, $Y_{ghstat}$, as the dependent variable. BMDP was initialized to commence with these variables in the model, including a constant term, and allow terms to move into or out of the model based upon the maximum likelihood method of selection of terms. The resulting model is

$$\log(odds) = \log\left(\frac{P(Y_{ghstat}=1)}{1-P(Y_{ghstat}=1)}\right) = .74192 + .10486 X_{sleep}$$

or

$$P(success) = P(Y_{ghstat}=1) = \frac{exp(.74192 + .10486 X_{sleep})}{1 + exp(.74192 + .10486 X_{sleep})}$$

Examination of the results shows that this model does not at all fit the data. With a p-value close to zero for the goodness-of-fit $\chi^2$ test statistic, the hypothesis of the model fitting the data is rejected. Plotting the number of hours of sleep against the natural log of the ratio of good to poor health status (figure 3.1), immediately reveals that one reason for this is that we are fitting a linear model to what is clearly not a linear phenomena.

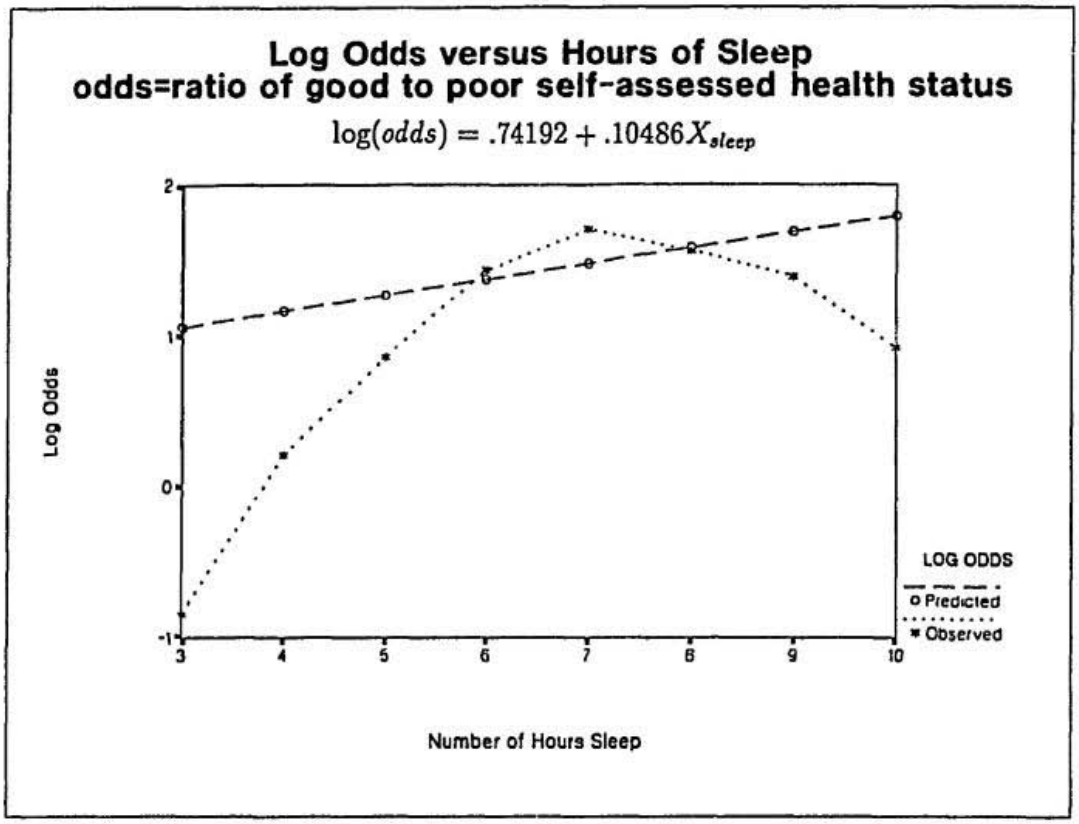


Figure 3.1: Log Odds versus Number of Hours Sleep (Linear Model)

Given the curved shape of the observed data, the program was run again, this time allowing for the independent variable $X_{sleep}$, and its square $X^2_{sleep}$ to enter into the model. That is, since the plot suggests a curvilinear relationship between odds of good to poor health and the independent variable sleep, a quadratic model in the variable sleep was fitted. The resulting model is

$$\log(odds) = -5.2570 + 1.8718X_{sleep} - .12654X^2_{sleep}$$

Based on the corresponding p-value of .901 we do not reject this model. The values predicted by the model are in close agreement with the observed data. Below we have a summary of selected results from this logistic regression program:

**Summary of Selected Results**

| Number of Hours Sleep | Number of Successes (Y=1) | Number of Failures (Y=0) | Observed Log Odds | Predicted Log Odds | Predicted Odds | Standard Residuals |
|---|---|---|---|---|---|---|
| 3 | 3 | 7 | -.8473 | -.7805 | .4582 | -.1124 |
| 4 | 16 | 13 | .2075 | .2055 | 1.2281 | .0069 |
| 5 | 80 | 34 | .8559 | .9384 | 2.5559 | -.4825 |
| 6 | 377 | 90 | 1.4326 | 1.4183 | 4.1301 | .1479 |
| 7 | 962 | 175 | 1.7043 | 1.6451 | 5.1815 | .9857 |
| 8 | 1060 | 223 | 1.5589 | 1.6188 | 5.0470 | -1.1852 |
| 9 | 140 | 35 | 1.3863 | 1.3394 | 3.8168 | .2890 |
| 10 | 50 | 20 | .9164 | .8070 | 2.2412 | .6091 |

The optimum average number of hours of sleep per night is 7 in that the predicted log odds are maximum at 7 hours sleep. At that point the predicted odds are 5.1815 (log odds= 1.6451). That is, under this model, given 7 hours sleep per night, the odds of reporting good health status (as opposed to poor) are 5 to 1. For those reporting only 3 hours sleep, the predicted odds of reporting good health status are .4582 to 1 (log odds= $-.7805$). In other words, these people are less likely – more than twofold – to say that they have good (versus poor) health. So then, people with 7 hours sleep are 11.3 times more likely than people with 3 hours of sleep to report good health; the odds ratio for people with 7 hours sleep versus 3 hours sleep is $\frac{5.1815}{.4582}$ or 11.3084. On the other end of the spectrum, those with 10 hours of sleep per night do not fare as badly as those with 3 or 4 hours per night but are worse, with respect to self-assessed health status, than those reporting 5 to 9 hours per night. The predicted odds ratio for those indicating an average of 10 hours sleep per night is 2.2412 to 1. This is clearly quite a bit less than the optimum ratio of 5.1815. The odds ratio comparing 7 hours sleep to 10 hours sleep is $\frac{5.1815}{2.2412} = 2.3119$ indicating that people with 7 hours sleep are 2.3 times more likely than people with 10 hours sleep to report a good health status versus a poor health status. This new model fits the data very nicely as is seen in figure 3.2.

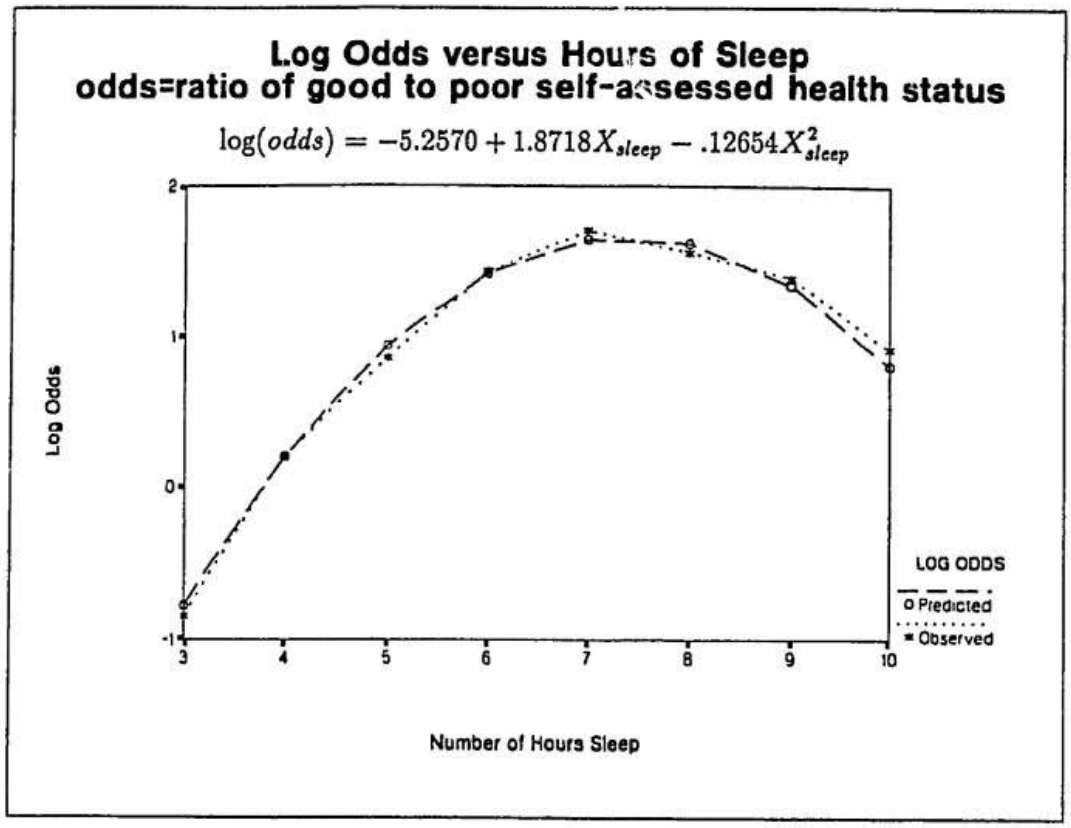


Figure 3.2: Log Odds versus Number of Hours Sleep (Quadratic Model)

The standardized residuals are very reasonable, fluctuating between $-1.1852$ and .9857 and showing no discernable pattern (figure 3.3). With the limited number of points, of course, any pattern might be difficult to perceive.

The probability plot of the predicted probability versus the observed proportion is very satisfactory (figure 3.4). As we would hope, our data are linear along a 45° angle. Plotting the predicted log odds against the observed proportion should result in a logistic curve. Given the few data points, it was difficult to claim this curve was definitely exhibited. The plot, however, was not unreasonable; it did not seem to deviate from a logistic trend.

Classification results for a variety of cutpoints are provided by BMDP. Using a

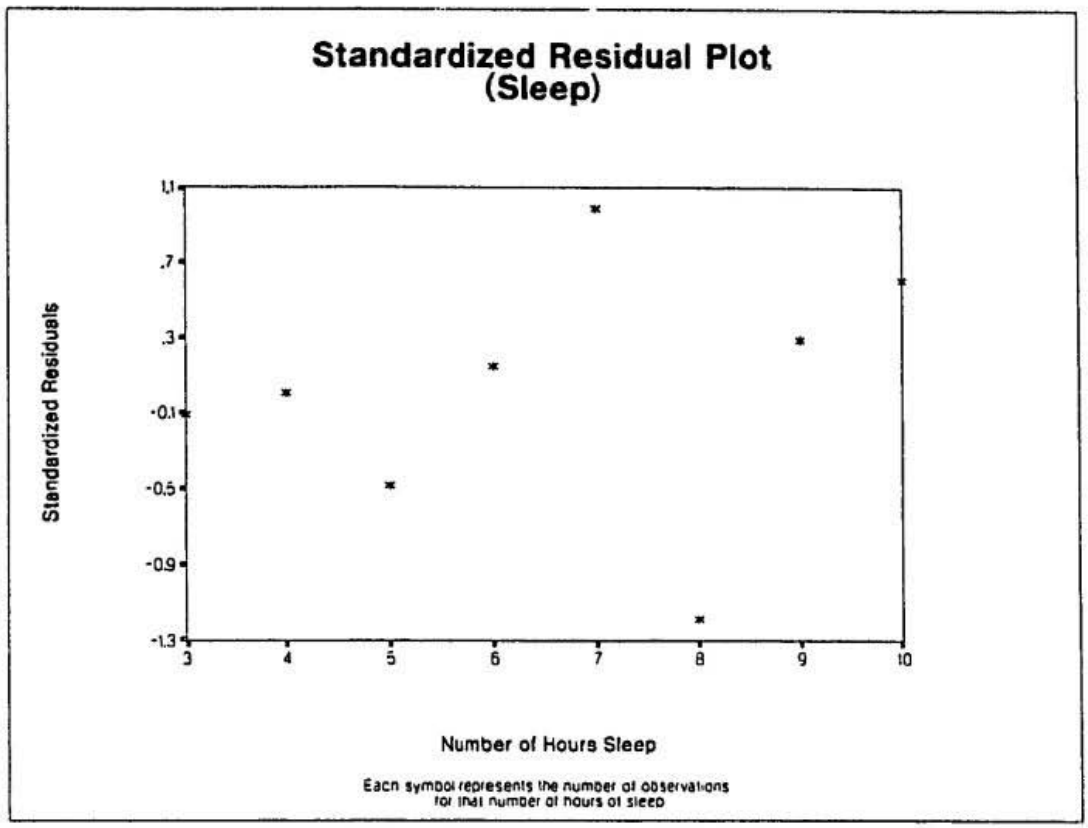


Figure 3.3: Standardized Residuals Plot (Sleep)

cutpoint of .813 gives the most satisfactory result overall. Using this cutpoint, based on our model a case will be said to belong to the group of those with poor health status if the predicted probability of success is $\leq .813$, and belong to the group of those with good health status if that predicted probability is $> .813$. At this cutpoint 75.22% of the successes (good health status) but only 33.33% of the failures (poor health status) are correctly predicted with an overall correct prediction of 67.16%. These classification results are not spectacular but nor are they startling. That is to say that although the data fits the model more than adequately, sleep alone is not sufficient for predicting the health status of an individual. Given the complex nature of health practices and their interactions with one another and with health

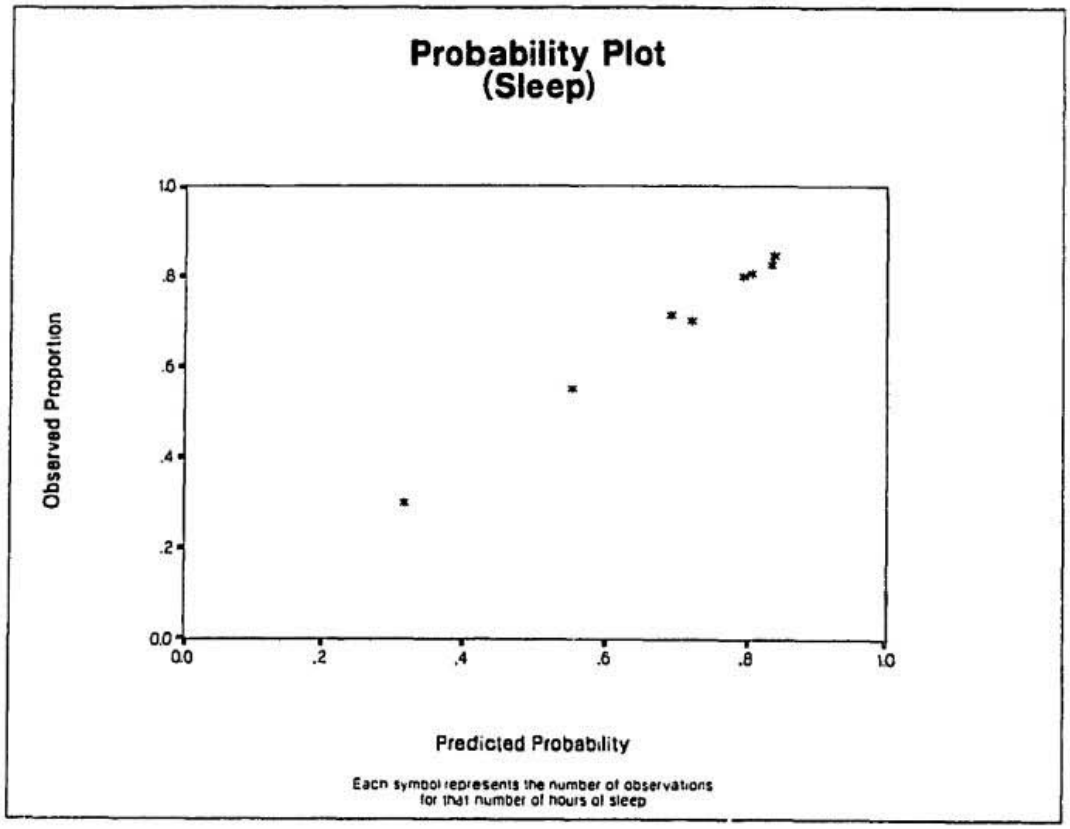


Figure 3.4: Probability Plot (Sleep)

status, this is not surprising. Nonetheless, our results are interesting and it remains that we can see from the logistic regression that the average number of hours sleep is associated with the binary health status variable with 7 hours being the optimum number of hours sleep for success of reported health status.

Note that there are other factors that the medical scientist might take into account to more fully describe the relationship between health and sleep. Age and sex are two such factors. So too is sleeping pattern. For instance, 7 consecutive hours of sleep might not have the same effect as 5 consecutive hours at night with a 2 hour afternoon nap each afternoon. As well, the effect of sleep may interact with daily activity levels and stress levels. For example, 7 hours sleep might have a different

effect on a person with a job requiring much physical activity but associated with low stress, than on a person with a stressful desk job requiring little physical activity. We can see then that even what appears to be a simple variable such as sleep may be quite complex and have complex interactions with other variables. This means that it is difficult to isolate and describe a pure sleep effect and to disentangle its effect on health status.

### 3.3.2 Drinking and Health Status

For some time it has been said that a moderate amount of drinking is not detrimental to one's health. In fact, it has been suggested that those who report having a drink a day also report having the best health. We study this relationship with our data set.

In looking at drinking behavior, we exclude those respondents who are not currently drinking but who did drink in the past. Although this subgroup are currently not consuming any alcohol at all, it is oversimplifying matters to include them in the same group as those who have never drank or who drink but very infrequently (less than once a month) as they differ somewhat from this group. Of the former drinkers, 64% said they had good health status. This is quite different from those who never drank of whom 78% claimed to have good health. The entire group of respondents excluding only the former drinkers, boasted almost 83% with good health. And we note that of the 127 former drinkers, 33% of them said they stopped drinking due to health reasons. Finally, we find that the distribution of drinking quantities differs substantially between former and non-former drinkers with the former drinkers consuming more alcohol when they did drink than the non-former drinkers currently consume.

In our first exploration of this variable's proposed association with health status, we use BMDP logistic regression with $X_{drink}$ as the independent variable and $Y_{ghstat}$ as the response variable. As was the case when we studied the relationship between sleeping and health status, we find that a linear function is not sufficient to describe the relationship between these variables. An examination of the plot of the number of drinks against the observed odds of good to poor health, suggests that it would again be appropriate to provide a quadratic term for possible inclusion into the model. When a quadratic term is included, the outcome is more successful resulting in a non-rejection of the hypothesis that the model fits the data (p-value=.849). The model is given as

$$\log(odds) = 1.3167 + .34536X_{drink} - .04875X^2_{drink}$$

When the former drinkers are included in this analysis as current non-drinkers, the overall trend is almost identical to what it is with them excluded and although the corresponding generated model does not fit the data quite as well (p-value=.677), it certainly fits adequately. This would in part be due to the fact that the former drinkers only made up for less than 4% of the total surveyed group so even though their behavior is different from other segments of the population (as outlined above), they constituted such a small number that the results would not have been unduly confounded had they been kept in the data set coded as current non-drinkers.

The optimum number of drinks per week is between 4 and 5 (figure 3.5). This quantity corresponds to a predicted log odds of good to poor health status of 1.9181 (odds=6.8080). In other words, those drinking 4 to 5 alcoholic beverages a week claim to have good health almost 7 times more frequently than poor health. This ratio decreases somewhat as the number of drinks decreases with those not drinking at all or drinking very infrequently boasting good health almost 4 times as often as poor

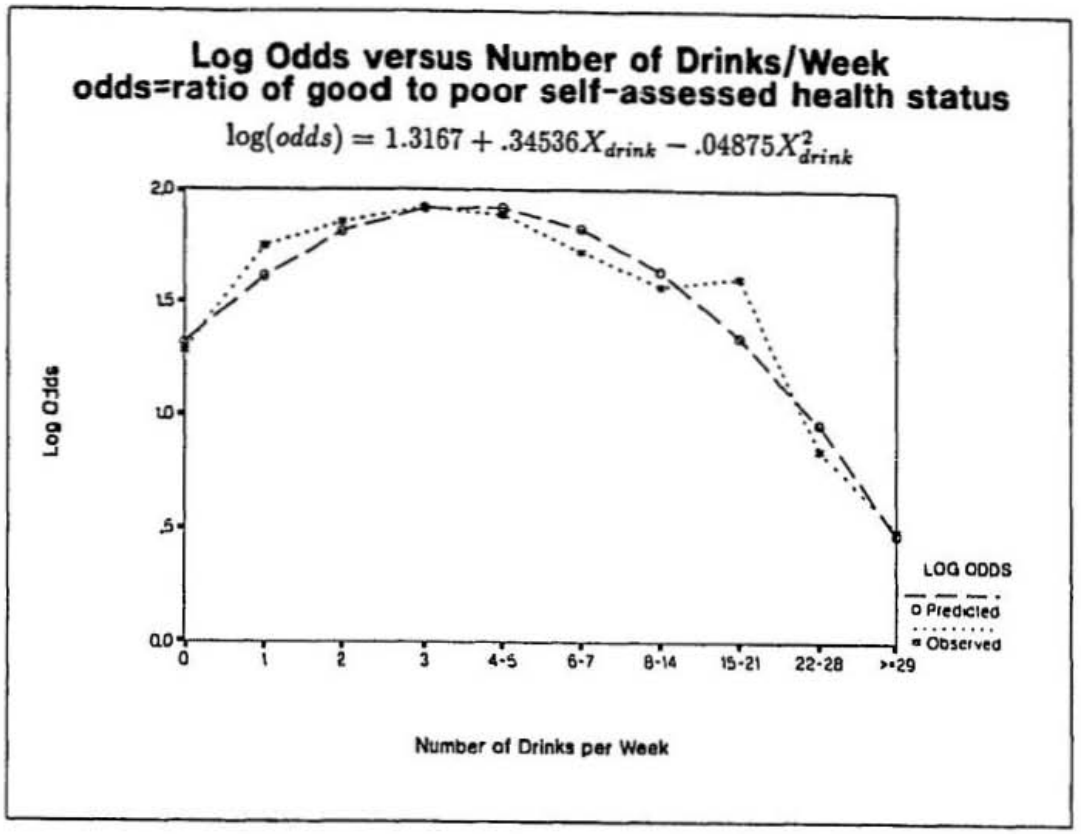


Figure 3.5: Log Odds versus Number of Drinks per Week

health (odds=3.7311). The odds decrease more as the number of drinks increases. The worst group consists of those who consume at least 29 drinks a week. That is, the group with the least favourable claim to good health average more than 4 drinks a day. For these people the predicted odds are 1.6095. So then, the odds of being in the good health category given one consumes the optimum of 4 to 5 drinks per week are $\frac{6.8080}{1.6095}$ or 4.2 times higher than the odds of being in the good health category given one consumes at least 29 drinks per week. That is, people are four times more likely to report good health if they drink moderately than if they drink excessively. The odds of reporting good health are $\frac{6.8080}{3.7311}$ or 1.8 times higher for those drinking 4 to 5 alcoholic beverages a week than for those not drinking at all. This supports previous

claims that moderate drinking (although not as much as a drink per day) is most often associated with good health. Although each category of drinking sees more people stating they have good rather than poor health, the ratio changes depending upon which category a person belongs to with the odds maximized for those drinking moderately and minimized for those drinking excessively.

The standardized residuals are acceptable, ranging from $-1.2365$ to 1.2413. The probability plot (figure 3.6) is also reasonable.

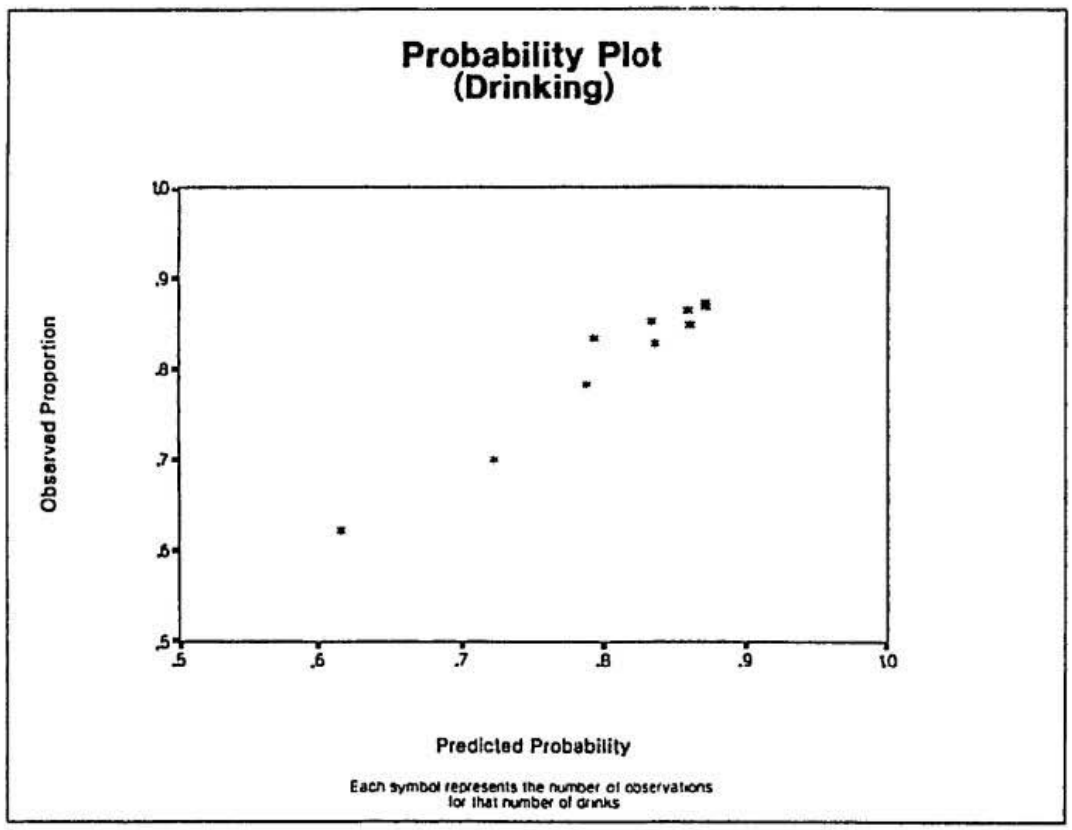


Figure 3.6: Probability Plot (Drinking)

The best we can do for classification results is to use a cutoff point of .813 which gives a 63.60% correct classification of successes and a 48.74% correct classification of failures for an overall correct classification rate of 61%. As before, this means that

one is categorized as belonging to the good or poor health status group depending on whether the predicted probability of success is $> .813$ or $\leq .813$, respectively. The percent correctly classified is not exceedingly high but the same rationalization exists here as with the health practice, sleep, namely that although a relationship clearly exists between drinking behavior and health status, this variable alone is not sufficient for predicting health status.

### 3.3.3 Drinking and Health Status, Controlling for Education

Health status is related to the amount of alcohol consumed. We speculate that patterns of drinking might change with educational level and hypothesize that health status improves with an increase in educational level. To see how alcoholic consumption and educational level interact and influence health status, further examination of these variables is required.

The logistic regression program is run as before but this time the variable education is factored into the equation. A person's educational level is categorized as being either at most high school (*educ*=1), post-secondary but not university (*educ*=2), or at least some university (*educ*=3).

With education, drinking and the square of drinking considered for inclusion into the model, the hierarchical rule was followed. That is, at any point no term may be in the model unless all its lower order terms, including main effects, are also in the model. The model generated by BMDP LR follows:

$$\begin{aligned}\log(odds) &= 1.8025 + .15275X_{drink} - .029506X^2_{drink} + .043818X(1)_{educ} \\ &\quad + .89772X(2)_{educ} + .039198X_{drink}X(1)_{educ} - .11140X_{drink}X(2)_{educ}\end{aligned}$$

where the odds are the ratio of good to poor health status. The p-value from the

above model is .362 so that the hypothesis of the model fitting the data is not rejected.

The standardized residuals, depicted in figure 3.7, range from −2.2131 to 1.9219, although most are between −1 and 1. The data in the probability plot (figure 3.8) follow a reasonably linear trend along a 45° angle.

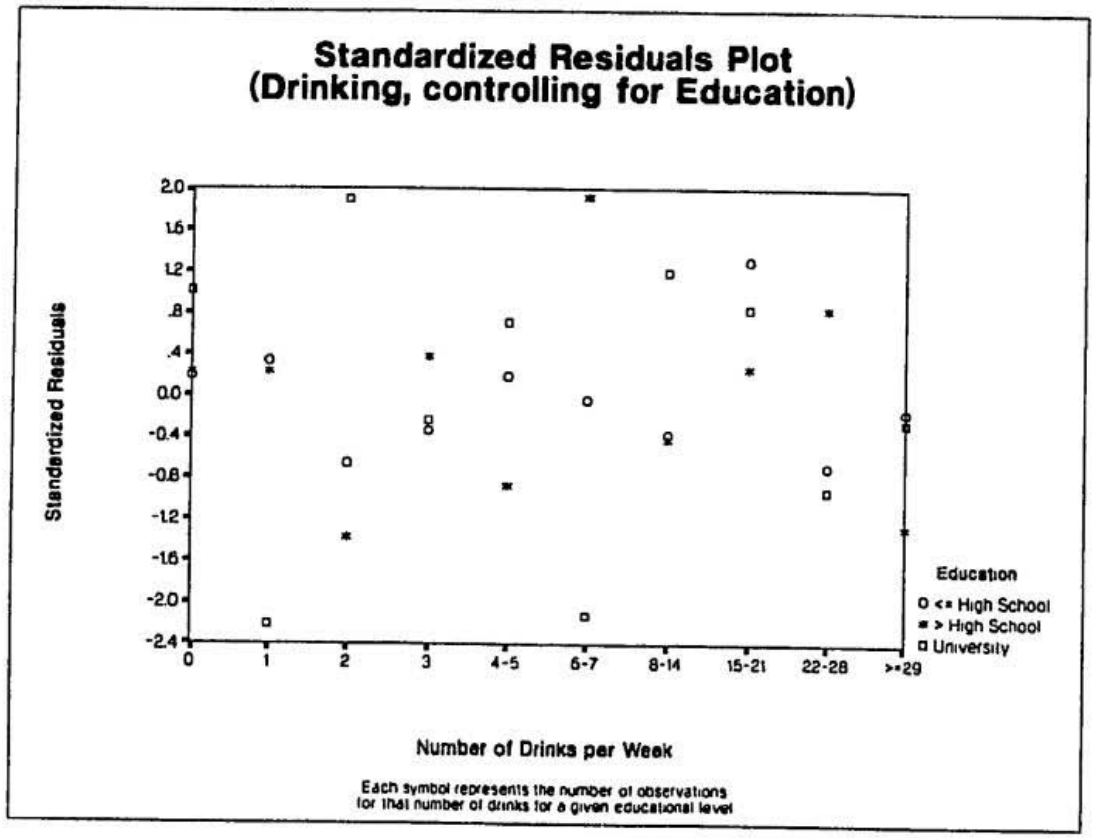


Figure 3.7: Standardized Residuals Plot (Drinking, controlling for Education)

The best cutoff point the model provides is between .788 and .813 when our percent correct classifications are 61.36% for success, 67.45% for failure and 62.42% overall. When we considered the relationship between health status and drinking, we found that although there is an association, our drinking variable alone was not sufficient for predicting one's health. When we control for education we discover that we cannot improve upon the power of prediction. While an association exists

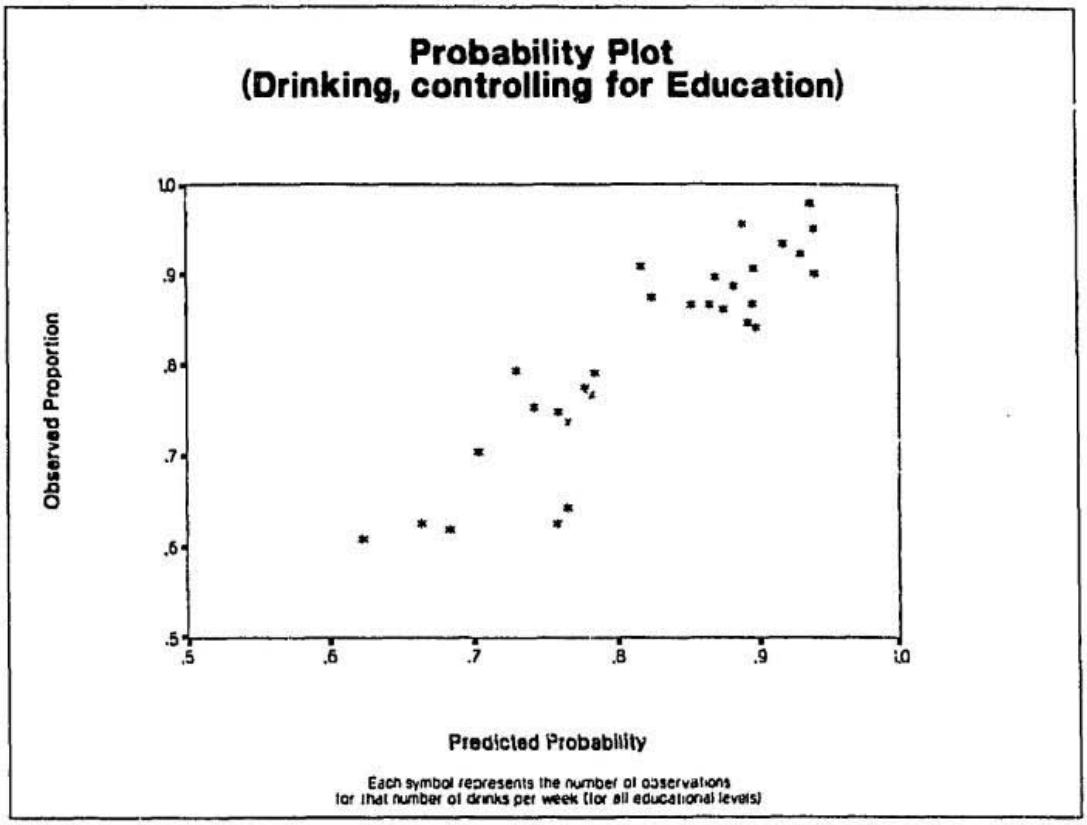


Figure 3.8: Probability Plot (Drinking, controlling for Education)

between drinking, education and health status, it remains clear that we are dealing with a complex phenomenon. Knowledge of both one's drinking behavior and educational background is still not sufficient for us to be able to predict one's self-assessed health status with any great degree of confidence. Including more variables might improve our power of prediction but then too our model will become more complicated. Intuitively this is what we might expect; we acknowledge that there are many behaviors and interactions between behaviors which will influence our overall health and well-being.

Regardless of the inability of the model to predict health status with a great degree of confidence, there are some interesting things to be gathered from it. Let us

examine it in more detail.

The design variables, $X(1)_{educ}$ and $X(2)_{educ}$ as seen in our model, take the values of $-1$ and $-1$ at the lowest educational level, *educ*=1. They take values of 1 and 0, respectively if *educ*=2, and values of 0 and 1, respectively at the educational level of at least some university, *educ*=3. Substituting these into the model produces the following three equations:

$$\begin{aligned}
educ{=}1 \quad \log(odds) &= .860962 + .224952X_{drink} - .029506X^2_{drink} \\
educ{=}2 \quad \log(odds) &= 1.846318 + .191948X_{drink} - .029506X^2_{drink} \\
educ{=}3 \quad \log(odds) &= 2.70022 + .04135X_{drink} - .029506X^2_{drink}
\end{aligned}$$

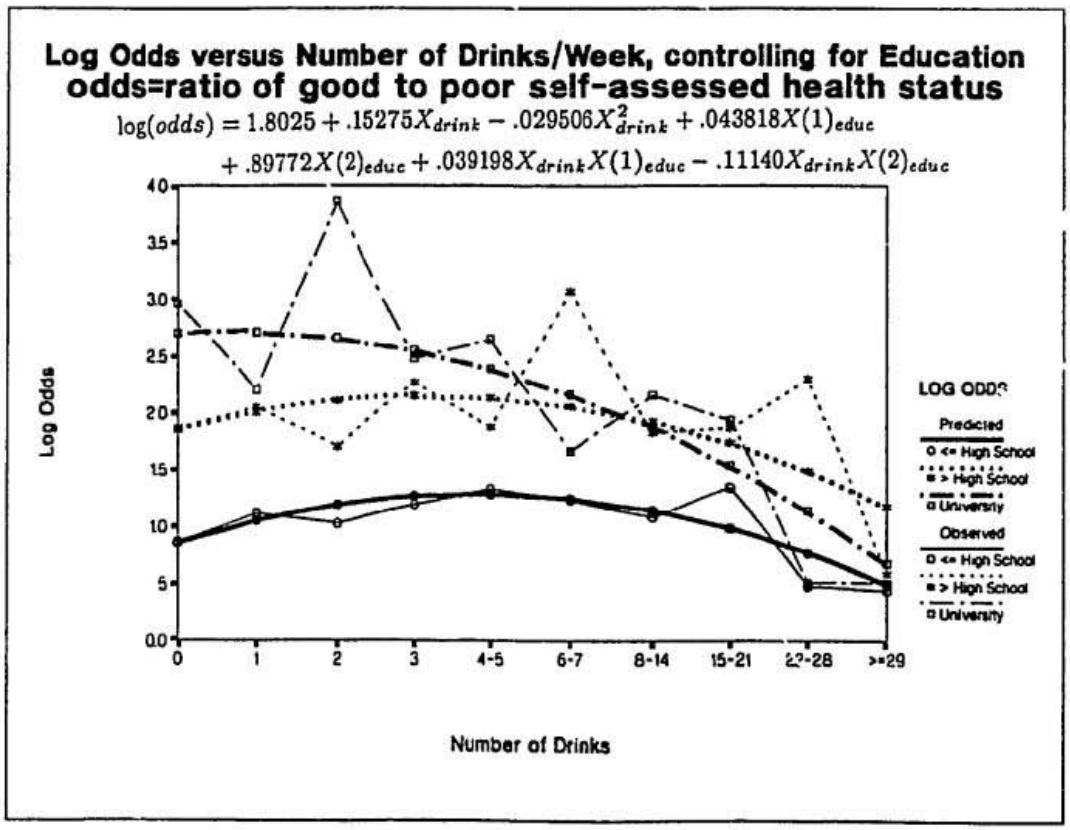


Figure 3.9: Log Odds versus Number of Drinks per Week, controlling for Education

In spite of the fact that the data do not fit the model as neatly as when education is not included, when we control for educational level a couple of interesting patterns are uncovered. In the first place, figure 3.9 immediately shows that regardless of the relationship between drinking and health status, in general one's health status improves as one's educational level increases. Secondly, the relationship between drinking and health status is extremely similar for the two lowest educational levels but this relationship differs markedly from that observed for the highest educational level.

When we look at the group having at most a high school diploma, we see that the ratio of good to poor health status peaks at 4 to 5 drinks per week and is worst at 29 or more drinks per week. This was the same as with the original model which did not factor education into the equation; however, the difference between the maximum and minimum odds ratios is not nearly as pronounced as when all educational levels were taken together. In both cases, at alcoholic consumption levels of at least 29 drinks per week, the odds of reporting good to poor health were less than 2 to 1 (odds=1.6). On the other hand, at the level of 4 to 5 drinks, when all education groups were taken together the odds were almost 7 to 1; this drops to under 4 to 1 for the same drinking category for the lowest education group.

The pattern for those in the next educational level is almost parallel to those belonging to the lowest level with the odds being greater at each drinking category for the higher educational group. It should be noted that the peak is now 3 drinks (odds=8.6) instead of 4 to 5 (odds=8.5) although the difference is barely discernable. This is the reverse of the lowest educational group, but for both educational levels the two categories of drinking with the best reports of health status (3, and 4 to 5 drinks per week) were extremely close.

The curve for the highest educational level is more dramatic than those of the two lowest levels. The group of people belonging to this highest level displays a different relationship between drinking and health. Good health is reported most frequently for those drinking one or no alcoholic beverage a week and the frequency steadily declines thereafter. As with the other groups, the greatest amount of alcoholic consumption corresponds to the minimum odds ratio of good to poor health. At this point, people report good health only twice as often as they report bad health. For these heaviest drinkers, this is a better ratio than for the heaviest drinkers with a maximum of high school but it is not as good as for the heaviest drinkers in the middle education group. For people with at least some university, the dramatic increase in good health status reporting belongs to those at the other end of the spectrum where having one or no drinks per week corresponds to reporting good health 15 times more frequently than reporting poor health. The predicted odds are 15.1 and 14.9, respectively, for these two drinking classifications. Although the ratio values drop off after this, they do not drop below even 9 to 1 until the drinking category increases to 6 or 7 drinks per week. Even then, we can continue to state that people in the highest educational group have better health than those in either of the two lower groups at almost every drinking level. Also, while a moderate amount of drinking is associated with reporting good health most frequently in the two lower educational groups, for those in the highest education group, consuming only one alcoholic beverage a week or not drinking at all is associated with the most frequently reported good health.

## 3.4 Hospital Utilization: A Loglinear Analysis

We have considered several models which look at the relationship between health status and health habit indicators such as sleeping and drinking practices. We now continue to examine health habit indicator variables further. A primary purpose of this study was to look not only at health practices but to examine how they relate to medical care utilization as measured by hospitalizations and visits to the doctor. In what follows, we wish specifically to see if practising good health habits is associated with whether or not one is hospitalized. We first develop models involving health habit scores and hospitalizations. From there we wish to develop a model considering the additional variables of sex, age, and education.

As described in chapter 2, the data were linked with hospital utilization data accumulated for the previous four years. This variable is coded as 0 or 1, where 0 is no hospital days for the previous four years and 1 is one or more hospital days in that same period. Hospitalizations due to pregnancy or delivery were ignored. The time frame of four years for hospital utilization data clearly differs from the snapshot in time of health practices as elicited from our sample in the questionnaire. Thus, in our analysis we assume that the health habits of the respondents at the point of the study are the same habits that they would have had for the previous four years. A currently on-going longitudinal study will track reported health practices and enable such assumptions to be validated.

### 3.4.1 The Health Practice Score

A considerable amount of effort went into creating an index of health practices. The following health practices – or lack thereof – are considered standard: eating breakfast, smoking, drinking, sleeping, correctness of weight, and exercising (see Belloc

and Breslow 1972, Belloc 1973, Breslow and Enstrom 1980, and Segovia et al. 1987).

After various preliminary exploratory analysis, such as the examination of measures of association, the breakfast variable – eating breakfast every day, occasionally or never – was dropped as a health practice. Its association with health status was negligible and consequently it was not included in the composite index of health practices. A person was considered to have a good health practice with respect to weight if he had correct weight as measured by the Quetelet index (Metropolitan Life Tables) where the Quetelet value is calculated as a function of a person's weight and height. A score of a moderate to very active exercise habit was coded as a good health practice as was an average of 7 to 8 hours sleep per night.

Some of the complexities inherent in gauging whether or not a person has a good drinking habit were discussed in the previous section. All things considered, a person is here accepted as having a good drinking habit if he or she has an average consumption of at most 5 alcoholic beverages per week.

It is also ambitious to try to categorize smoking habits as simply good or bad. Obviously it is a good if an individual never smoked. So too it is not good if an individual does smoke, although the degree of 'badness' changes substantially depending upon the amount smoked and the duration of the habit. Former smokers are much more difficult to classify as having a good or bad smoking habit. Much is written that acknowledges that smoking cessation is unquestionably good. However, the amount of time required before an ex-smoker approaches the same risk level as 'never-smokers' of diseases known to be worsened by smoking – such as heart disease and lung cancer – is dependent upon such factors as the length of time since cessation, the amount of smoking while a smoker, and whether cessation occurred after the onset of smoking-related diseases. Although the benefits of cessation are almost

immediate, the literature suggests that dependent upon the aforementioned factors, an ex-smoker does not approach the same risk level as 'never-smokers' until after 5, 10 or even 15 or more years. For this reason, for the purpose of our analysis we only count those who have never smoked as having a good smoking health practice. For discussions of smoking and consequences of cessation, see Cook et al. (1986), Griffith and Garcia (1989), Warner (1989), Belt (1990), Miller et al. (1990), and the U.S. Department of Health and Human Services (1990).

Although each health practice was coded as either good (1) or bad (0), they do not carry equal importance as health habits. The frequency distribution of each variable was studied and each one was cross-tabulated against self-assessed health status to provide measures of association to assist in the assigning of appropriate weights for a health practice score. This, together with an examination of logistic analysis led to the following weights:

| Weighting Factor | Positive Health Practice |
|---|---|
| 4 | smoke (never) |
| 4 | exercise (moderately to very active) |
| 3 | weight (correct - Quetelet index) |
| 2 | drink (maximum of 5 per week) |
| 2 | sleep (7 to 8 hours per night) |

Other weights have been studied with this data set (Segovia et al. 1987) but given the exploratory analysis, the above is a reasonable weighting distribution for a health practice score. This score ranges from 0, when all health habits are negative, to 15 when they are all positive. We might express the health practice score as an equation with weights,

$$\textit{Health Practice Score} = \sum_i w_i p_i$$

where $w_i$ is the weight of the $i^{\text{th}}$ health practice and $p_i$ is an indicator variable for the $i^{\text{th}}$ health practice, taking a value of 1 if the health habit is practised and 0 if

it is not. The scores calculated from each possible combination of health practices were studied after which a grouped health practice score of three levels was created. Scores of 0 to 7 belong to those with the lowest level of positive health habits. This is followed by those with scores between 8 and 10. The group with scores between 11 and 15 have the highest level on this grouped health practice score. We may note at this point that we could look at logistic regression using the full health practice score from 0 to 15 rather than a grouped score. This, in fact, was examined briefly before continuing with a loglinear analysis.

### 3.4.2 Modelling for Hospitalizations and Grouped Health Practices

In trying to determine the relationship between the dichotomous hospitalization variable and the health practices information, various models were fitted to the data using the procedure, LOGLINEAR in SPSS-X. Although it is common to treat all categorical variables as nominal, the grouped score for health practices is an ordinal variable. Agresti (1984) discusses how one might take this ordinality into account by testing a somewhat more complex model than that of simple independence of the two variables.

With the hospitalization observations as the nominal row variable $H$, and the grouped health practice score as the ordinal column variable $P$, we test the row effects model

$$\log F_{ij} = \mu + \lambda_i^H + \lambda_j^P + \tau_i(v_j - \overline{v})$$

which uses the standard notation where $\mu$ is the grand mean of the logs of the expected frequencies and $\lambda_i^H$ and $\lambda_j^P$ are the terms for the main effects of hospitalization and the grouped health practice score, respectively. In the last term, $v_j$ is the $j^{\text{th}}$ score of the column variable $P$, and $\tau_i$ is the slope for row $i$ $(i = 1, 2)$ of the deviation within that row, of $\log F_{ij}$ from the simple independence model, $\log F_{ij} = \mu + \lambda_i^H + \lambda_j^P$.

Furthermore, $\sum \lambda_i^H = \sum \lambda_j^P = \sum \tau_i = 0$.

With a likelihood ratio $G^2 = 0.094$ with 1 degree of freedom and a corresponding p-value=.759, we do not reject the hypothesis that the row effects model is a good fit. That is, the model fits the data well when the grouped health practice score is treated as ordinal.

When the simpler model of independence which ignores the row effects term is fit to the data, it gives a likelihood ratio value of $G^2$=2.640 with 2 degrees of freedom and a p-value of .267. Again we do not reject our hypothesis that the model fits the data well.

For the test of the hypothesis of independence given row effects, we calculate $G^2$ as the difference in $G^2$s between the independence and row effects models. This produces $G^2$=2.546 with 1 df which leads to a non-rejection of the hypothesis; given that the row effects model is satisfactory, we additionally claim that health status and health practices are not associated when the ordinality of the health practices score is considered.

In summary, we partition $G^2$ to give the following table:

| | Model | df | $G^2$ |
|---|---|---|---|
| Independence model | $\log F_{ij} = \mu + \lambda_i^H + \lambda_j^P$ | 2 | 2.640 |
| Row effects model | $\log F_{ij} = \mu + \lambda_i^H + \lambda_j^P + \tau_i(v_j - \overline{v})$ | 1 | .094 |
| Independence, given row effects | | 1 | 2.546 |

It is reasonable not to reject whichever of these models of independence we adopt. As is generally the case under such circumstances, we fit the simpler model. The simpler independence model implies that hospitalization is independent of health practices. It is commonly understood from medical studies, however, that good health

practices can improve one's health and longevity. We would thus strongly expect health practices to be related to hospitalizations. The fact that this relationship does not appear here may not necessarily be due to the absence of the existence of such an association. Variables such as health practices and hospital utilization are complex and may not be captured by these models and, in particular, by our current variable measures.

### 3.4.3 Examination of the Row Effects Model

Although we fit the independence model, the row effects model was not rejected so it is interesting to briefly study this model a little further. Let us examine our contingency table in terms of the row effects model.

| | Grouped Health Practices (P) | | | |
|---|---|---|---|---|
| Hospital (H) | 1=low (0-7) | 2=medium (8-10) | 3=high (11-15) | Totals |
| 0 days | 1117<br>(1115.59)<br>[1130.69] | 667<br>(669.83)<br>[666.16] | 504<br>(502.59)<br>[491.15] | 2288 |
| ≥1 day | 285<br>(286.41)<br>[271.31] | 159<br>(156.17)<br>[159.84] | 105<br>(106.71)<br>[117.85] | 549 |
| Totals | 1402 | 826 | 609 | 2837 |

The first entry of a given cell is the observed frequency, the second is the expected frequency under the row effects model and the third is the expected frequency under the independence model.

From running the SPSS-X loglinear program for the row effects model, we have $\hat{\tau}_1 = .048$, and since $\sum_i \hat{\tau} = 0$, it follows that $\hat{\tau}_2 = -.048$. While the slopes are close to zero, their direction is as we would expect. For the first row ($H$=0) $\hat{\tau}_1$ is positive, while for the second row ($H$=1) $\hat{\tau}_2$ is negative. Since $\hat{\tau}_1 > \hat{\tau}_2$, of the two hospitalization groups, the first group is the one with the better health habits where a better health habit is reflected in a higher grouped health practice score if we accept

the ordinal nature of this variable. It is not surprising that this positive trend is exhibited by the group without any hospital days. On the other hand, the negative slope for the second group reflects the decreasing probability of at least one hospital day as the health practice score increases.

Put another way, with the row fixed at $i = 1$, the positive slope $\hat{\tau}_1$ reflects the increased probability of no hospital days as the number of health practices increases. With the row fixed at $i = 2$, the negative slope $\hat{\tau}_2$ indicates that the probability of having at least one hospital day increases as the health habits index decreases.

If we think of this in terms of odds ratios, under the row effects model we say that conditional on belonging to the first row - that is, having no hospital days - the odds of having a low score ($P$=1) as opposed to a medium score ($P$=2) are $\frac{1115.59}{669.83} = 1.67$. The odds change only slightly to $\frac{286.41}{156.17} = 1.83$ for those belonging to the group having at least one hospital day. The odds ratio, therefore, is $\frac{1.67}{1.83} = 0.91$, so that the odds of having a low number of health practices rather than a medium number, are almost the same for the group with no hospital days as for the group with one or more days in hospital. The same value for the odds ratio exists for the other two adjacent columns - that is when comparing hospitalizations for those with a medium health practice score to those with a high score. This ratio is very close to 1 and is even closer to unity under the simple independence model which ignores the ordinality term.

The odds of having a low rather than a high health practice score are 2.22 and 2.68 for $H$=0 and $H$=1, respectively. The resulting odds ratio of .83 is still close to unity although not as close as when we examined adjacent columns. Although the odds ratio for the adjacent columns were the same, the magnitude of the ratio changed for the extreme columns due to the ordinality of the column variable. Under

the simple independence model which ignores the ordinality, the odds ratio for the extreme columns is the same as the ratios for the adjacent columns, that odds ratio being essentially unity.

The odds ratios for the adjacent columns can also be computed directly from the slopes $\tau_i$, $i = 1,2$ since the difference between them is equal to the log of the odds ratio. And since $v_j = j$, the odds ratios for the adjacent columns are equivalent (see Agresti 1984). That is,

$$\hat{\tau}_2 - \hat{\tau}_1 = \log \frac{f_{11}f_{22}}{f_{12}f_{21}} = \log \frac{f_{12}f_{23}}{f_{13}f_{22}} = \log \hat{\alpha}$$

so that

$$\hat{\alpha} = e^{\hat{\tau}_2 - \hat{\tau}_1} = e^{-.096} = .91$$

We note here that the above contingency table is a small two-way table and the cell frequency counts are reasonably large. We therefore also looked at the relationship between health practices and hospitalizations without grouping the health practice score into three categories. Instead the health practice index was kept in its original form as a weighted score taking values from 0 to 15. It was then treated first as an ordinal and then as a nominal variable. The same conclusions were drawn however - namely of no association between health practices and hospitalization - which suggests that our grouping cutpoints into the three categories are reasonable. It is necessary to have such a grouping as we add variables to the model thus increasing the number of cells in the contingency table. Then, even with our reasonably large sample of approximately 3000 individuals, the number of variables involved requires that categories be collapsed in order that the expected frequencies in the cells of the contingency tables be acceptable. As the number of cells increases so too does the possibility of too many sampling zeros which makes the asymptotic sampling distribution assumptions invalid.

### 3.4.4 Fitting a more complex model

The model of independence was not rejected. When we test for association between health practices and hospitalizations, $G^2$=2.640 and df=2 for a p-value of .267. Hosmer (1989) mentions that if we are going to continue to add terms to a model, individual terms which when tested for significance give p-values of magnitude up to approximately .25 might not necessarily be immediately discarded as insignificant. A term may interact with other variables in such a way that it could remain in a more complex model as part of an interactive term. We wish to further examine our hospitalization variable when sex, age, and education are included in the model. The health practice score is left in as it may be associated with one or more of these variables as they in turn interact with the hospitalization variable.

The SPSS-X hierarchical loglinear program, HILOGLINEAR, with backward elimination was run. This program commences with the saturated model and deletes one term at a time until a simpler model is generated. As a first step the following variables were entered into the model: hospitalization ($H$), grouped health practice score ($P$), sex ($S$), age ($A$), and education ($E$). Age was grouped as 20-44, 45-64, or $\geq$65 years old. Education was coded into two levels - at most a high school education or at least some post-secondary education. Other cutpoints were tried with these two variables but the results were the same in as much as the same models were generated under these different groupings. The cutpoints given above then, appear acceptable.

The saturated model is given by

$$\log F_{ijklm} = \mu + \overbrace{\lambda_i^H + \lambda_j^P + \lambda_k^S + \lambda_l^A + \lambda_m^E}^{\text{main effects}} + \overbrace{\lambda_{ij}^{HP} + \lambda_{ik}^{HS} + \cdots + \lambda_{lm}^{AE}}^{\text{2-way interactions}} + \underbrace{\lambda_{ijk}^{HPS} + \lambda_{ijl}^{HPA} + \cdots + \lambda_{klm}^{SAE} + \lambda_{ijkl}^{HPSA} + \lambda_{ijkm}^{HPSE} + \cdots + \lambda_{jklm}^{PSAE} + \lambda_{ijklm}^{HPSAE}}_{\text{3,4,5-way interactions}}$$

which includes the fifth order interaction term and each two, three and four-way

interaction term as well as the main effects – hence the word 'hierarchical' to describe the model. A more convenient notation for this model is $[HPSAE]$ which is the five-way interaction term. All simpler interaction terms and main effects are implied by this notation.

The results of the tests that the k-way and higher order effects are zero for k=1,...,5 and that the k-way effects *only* are zero are given below:

| k-way and higher | | | |
|---|---|---|---|
| k | df | $G^2$ | p-value |
| 5 | 4 | 3.012 | .5558 |
| 4 | 20 | 29.283 | .0823 |
| 3 | 45 | 104.200 | .0000 |
| 2 | 64 | 663.776 | .0000 |
| 1 | 71 | 3730.029 | .0000 |

| k-way | | | |
|---|---|---|---|
| k | df | $G^2$ | p-value |
| 1 | 7 | 3066.252 | .0000 |
| 2 | 19 | 559.576 | .0000 |
| 3 | 25 | 74.917 | .0000 |
| 4 | 16 | 26.271 | .0503 |
| 5 | 4 | 3.012 | .5558 |

From the first table, we do not reject the hypothesis that the $5^{th}$ order effect is zero and at the 5% level of significance, we also do not reject the hypothesis that the $4^{th}$ and $5^{th}$ order interactions are zero.

In testing that the k-way effects are zero, we conclude that the $1^{st}$, $2^{nd}$ and $3^{rd}$ order effects should be added to the grand mean in the model. While the $5^{th}$ order effect need not be included, it is questionable whether we should include any $4^{th}$ order terms. The test that these effects are zero yield $G^2$=26.271, df=16 with p-value=.0503 so using *exactly* a 5% level of significance, we would choose not to include four-way interaction terms. It would be desirable if no $4^{th}$ order effects were included but if instead a simpler model, with at most three-way interaction terms, were to fit the data reasonably well.

The hierarchical model that is produced has generating class $[AE,PE,PSA,HSA]$ which, in recalling the hierarchical notation, means that all lower order terms are also in the model. With $G^2$=53.106, df=43 and p-value=.139, we do not reject the hy-

pothesis that this model is a good fit. Upon closer examination of the contribution of the two 3rd order terms to this model, it seems likely that as it exists here, both these higher order terms will need to remain in the model for it to fit the data well. This will be discussed presently.

There are associations in this model. It would be useful, therefore, to know the underlying nature of the dependency. One possibility is to model for linear dependency by considering the possible linear trends in the odds ratios due to the ordinality of the variables. While in our hierarchical model each variable is treated as nominal, the three levels for both age and grouped health practices are ordinal. To examine this more closely, we take the model as generated by the HILOGLINEAR program and run the LOGLINEAR program with the same terms specified but now treating the variables age and grouped health practice score as ordinal. The design specification is

$$\begin{aligned}\log F_{ijklm} = \ & \mu + \lambda_i^H + \lambda_j^P + \lambda_k^S + \lambda_l^A + \lambda_m^E + \lambda_{ik}^{HS} + \tau_i^{HA}(v_l - \overline{v}) + \tau_k^{PS}(u_j - \overline{u}) \\ & + \beta^{PA}(u_j - \overline{u})(v_l - \overline{v}) + \tau_m^{PE}(u_j - \overline{u}) + \tau_k^{SA}(v_l - \overline{v}) + \tau_m^{AE}(v_l - \overline{v}) \\ & + \tau_k^{PSA}(u_j - \overline{u})(v_l - \overline{v}) + \tau_{ik}^{HSA}(v_l - \overline{v})\end{aligned}$$

where

$$\begin{aligned}\sum \lambda_i^H \ &= \sum \lambda_j^P = \sum \lambda_k^S = \sum \lambda_l^A = \sum \lambda_m^E = \sum_i \lambda_{ik}^{HS} = \sum_k \lambda_{ik}^{HS} = \sum \tau_i^{HA} \\ &= \sum \tau_k^{PS} = \sum \tau_m^{PE} = \sum \tau_k^{SA} = \sum \tau_m^{AE} = \sum \tau_k^{PSA} = \sum_i \tau_{ik}^{HSA} = \sum_k \tau_{ik}^{HSA} = 0.\end{aligned}$$

Note that $i = 1, \ldots, h; j = 1, \ldots, p; k = 1, \ldots, s; l = 1, \ldots, a; m = 1, \ldots, c$, so that this model has the degrees of freedom given below. This is left unsimplified so as to indicate the degrees of freedom for each estimated parameter.

$$\begin{aligned}\text{df} \ &= hpsae - [1 + (h-1) + (p-1) + (s-1) + (a-1) + (e-1) + (h-1)(s-1) \\ &\quad + (h-1) + (s-1) + 1 + (e-1) + (s-1) + (e-1) + (s-1) + (h-1)(s-1)]\end{aligned}$$

$HS$ is an interaction term between two nominal variables. The association terms for $HA$, $PS$, $PE$, $SA$ and $AE$ are the different row effects similar to that which

was discussed earlier in the row effects model which had only the two variables, hospitalization and health practices. The $PA$ term is a linear-by-linear association for the two ordinal variables, grouped health practices and age. The $PSA$ effect is an association between these two ordinal variables with the nominal variable sex. Finally, $HSA$ is the association term for the nominal variables hospitalizations and sex, and the ordinal variable age.

This model gives a goodness-of-fit test statistic, $G^2$=86.77069 with df=55 and p-value=.004 so that the hypothesis that the model fits well is rejected. Whatever dependency is exhibited between the variables in the hierarchical model, it cannot be explained in terms of a linear trend by simply adding the ordinal effects of age and grouped health practices into the model. The hierarchical model was also adjusted to account for the row effects terms for age or health practice score alone while the other variable was treated as nominal. The models produced from this also led to rejecting the hypotheses of a good fit.

Since the ordinality of two of the variables does not explain the association in our model, we return to our original hierarchical model $[AE,PE,PSA,HSA]$ which was an acceptable fit. While it is a good fit, closer examination might suggest terms which could be dropped without seriously reducing the goodness-of-fit.

When the estimates for this model are examined, for instance, we see that for one of the four estimated parameters for the three-way interaction of health practice score by sex by age, the hypothesis that the parameter is zero is rejected at the 5% level of significance. Dropping this three-way interaction term while retaining the lower effects generated by it would simplify our model somewhat and make it more readily interpretable. For the partial association of this term in the saturated model, however, the observed significance level is very small for this association,

implying that it might be an important term to retain in the model. When this term was dropped from the hierarchical model, the resulting model was not a good fit so it was rejected. The results would be similar if the other three-way effects term, hospitalization by sex by age, were dropped while keeping its implied lower order effects. The remaining $Z$-values of the estimated parameters for the hierarchical model were such that all the corresponding terms were kept in the model.

As well as investigating the possibility of dropping terms from our model, there may be other terms which are not added by the HILOGLINEAR procedure but which we intuitively believe should be in the model. The interaction between hospitalizations and grouped health habits, $HP$, is one such term. While the only two-way interaction terms missing from this model are $HP$, $HE$ and $SE$, it is $HP$ which is the the most notable. For what we are studying, it may not be too interesting that such a model does not indicate a significant relationship between sex and education when all the other variables are considered. The lack of association between hospitalizations and education may be more interesting but not too surprising. We might well expect, however, that hospitalizations and health practices interact in such a way as to contribute significantly to the fit of the model. In our earlier independence and independent row effects models, however, this was not the case. This lack of association between these two variables was reinforced again in this more complex hierarchical model so we do not add the term to the model. Our final model, therefore, is the one which was generated by the HILOGLINEAR procedure, namely $[AE,PE,PSA,HSA]$ or

$$\begin{aligned}\log F_{ijklm} &= \mu + \lambda_i^H + \lambda_j^P + \lambda_k^S + \lambda_l^A + \lambda_m^E + \lambda_{ik}^{HS} + \lambda_{il}^{HA} + \lambda_{jk}^{PS} \\ &\quad + \lambda_{jl}^{PA} + \lambda_{jm}^{PE} + \lambda_{kl}^{SA} + \lambda_{lm}^{AE} + \lambda_{jkl}^{PSA} + \lambda_{ikl}^{HSA}\end{aligned}$$

This indicates three-way interactions between our main variable of interest, hospital utilization, and sex and age, and between health practices, sex and age, together

with the two-way interactions stemming from these associations. The other two-way interactions which exist are between health practices and education, and age and education. These associations will be discussed presently.

### 3.4.5 Residuals for the Hierarchical Model [*AE*,*PE*,*PSA*,*HSA*]

It remains to examine the residuals. Of our adjusted standardized residuals for the 72 cells, only three exceed 1.96 with values of 2.05, 2.30 and 2.41. This is reasonable since we would by chance expect 5% of the residuals to exceed 1.96. Of the remaining residuals, all but nine do not exceed 1.645 in absolute value.

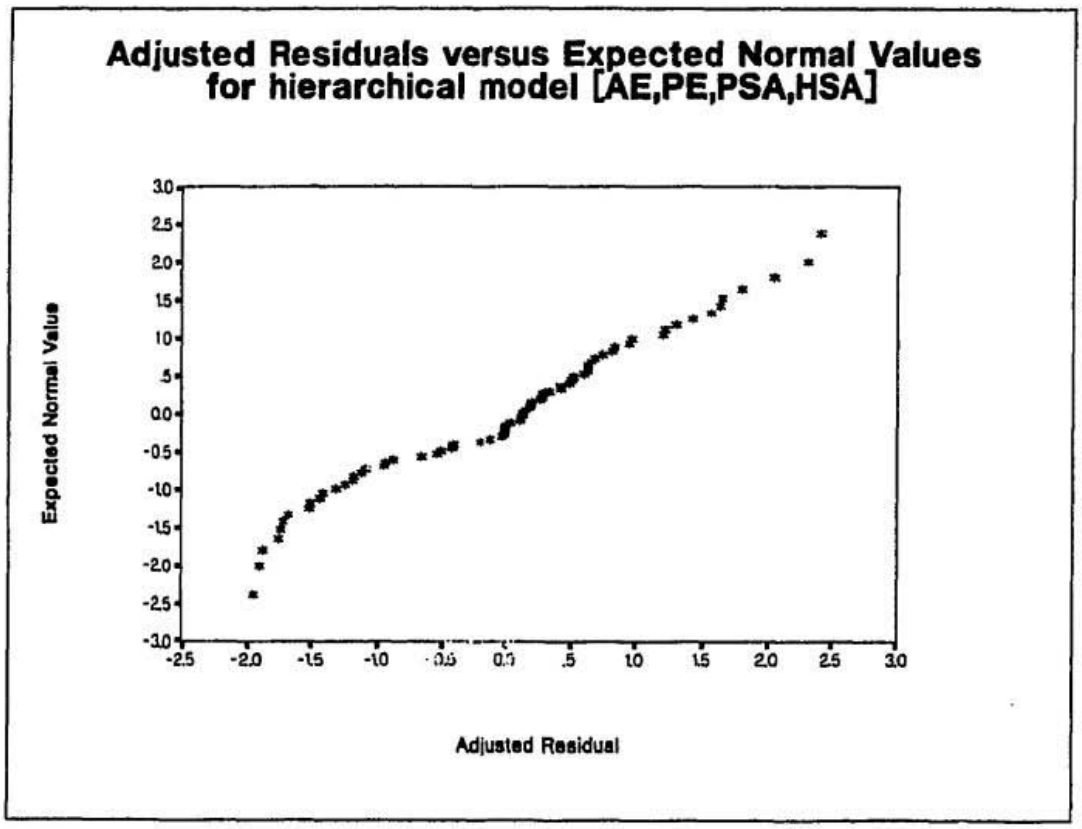


Figure 3.10: Adjusted Residuals versus Expected Normal Values, [*AE*,*PE*,*PSA*,*HSA*]

The normal plot of the adjusted residuals against the expected normal values (figure 3.10) is almost linear along the diagonal with the small deviation that does exist being predominantly for negative residuals. The correlation between the adjusted residuals and their expected normal values is .988. A correlation test for normality, equivalent to the Shapiro-Wilk test (Minitab Reference Manual 1989), results in a non-rejection of the hypothesis that the adjusted residuals are normally distributed.

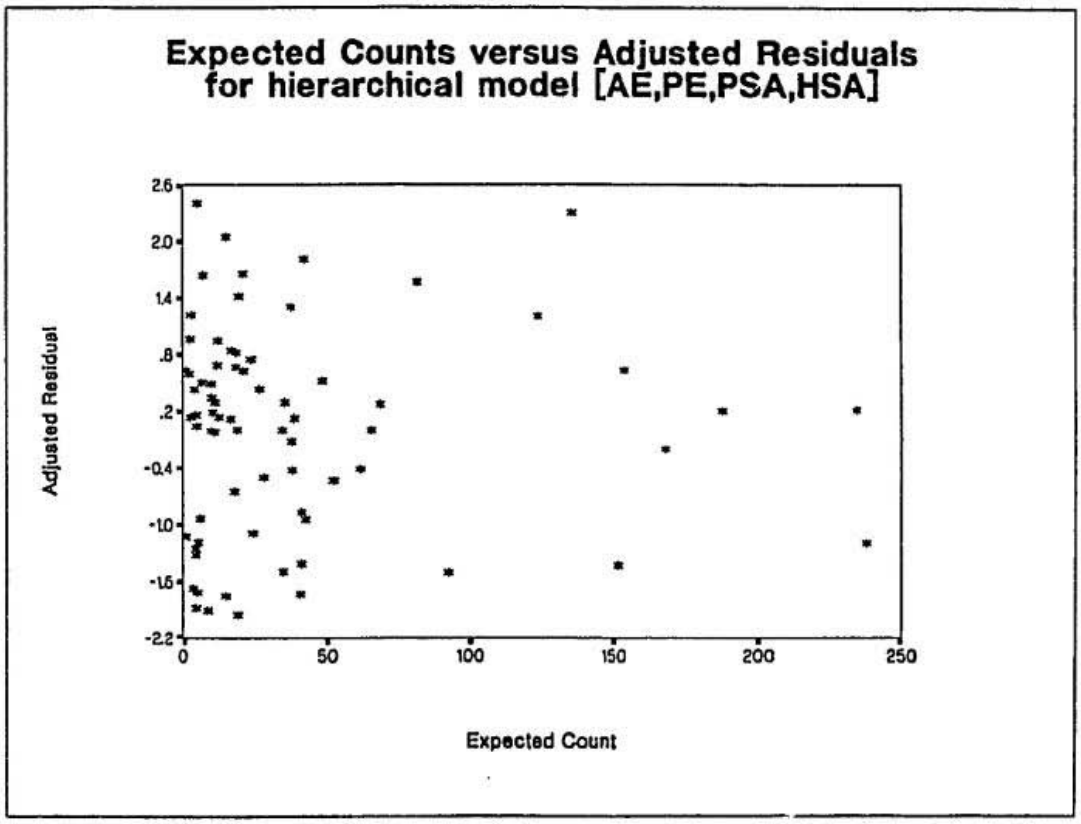


Figure 3.11: Expected Counts versus Adjusted Residuals, [*AE*,*PE*,*PSA*,*HSA*]

The plot of the expected counts against the adjusted residuals is shown in figure 3.11. There is a slightly discernable pattern displayed here showing that the comparatively largest residuals correspond with those cells for which the expected

counts are comparatively smallest. This suggests that the model does not fit as well for cells with smaller expected frequencies as it does for cells with larger expected frequencies.

### 3.4.6 Summary

Our hierarchical model is a reasonable one. All the main effects are included in this model. As well, most of the two-way interaction effects are there. The three-way interaction of grouped health practice with sex and age ($PSA$) is such that the strength of association between health practices and age is somewhat greater for males than for females. First we note that for males the value of Cramér's $\hat{V}$ is .115 while for females it is .083. Although both values are very low, we note that while Cramér's $V$ is not a useful measure of association in and of itself, it can be useful in comparing the magnitude of association across several tables. Hence we can say that it appears that the strength of association between grouped health practices and age is greater for males than for females. Since our marginal distributions between the two tables are not dissimilar, we have some additional degree of confidence in this measure. In spite of this, we would not rely upon this as a sole measure of relationship.

With both grouped health practices and age being ordinal variables, we use $\gamma$ as one measure to test the strength of the association between these variables, controlling for sex. For males, $\hat{\gamma} = -.274$ and for females $\hat{\gamma} = .150$. Neither of these values is strong and in addition we recall that $\gamma$ tends to exaggerate the true association. Nonetheless the weak relationship is interesting to look at in terms of the difference in the direction of association of the sexes. Recall from the earlier discussion of measures of associations that one interpretation of $\gamma$ is in terms of the difference in proportion of concordant and discordant pairs of individuals. In the

case of males, $\gamma$ is negative so that the number of discordant pairs is greater than the number of concordant pairs. That is, there are more pairs for which if one male rates higher than the other male on grouped health practices or age, then he rates lower on the other variable. For females, on the other hand, $\gamma$ is positive so that there are more concordant than discordant pairs. In other words, there are more pairs for which one female ranks higher than the other female on both grouped health practices and age. Although weakly associated, it seems that while men become less active as they age, women become more active. Looking at this in terms of a PRE interpretation, we recall that $\gamma$ is the proportional reduction in error in predicting the order of pairs when we move from having no knowledge of their order, to using the knowledge of the number of concordant and discordant pairs to guess the order of each pair. As pointed out by Mueller et al. (1977), the signs show that we should use our knowledge of the number of concordant and discordant pairs to guess discordance for males and concordance for females when looking at grouped health practices with age. The proportional reduction in predictive error is 27.4% and 15.0% for males and females, respectively.

We examine the $PSA$ interaction a little further. Let us look at the relationship between grouped health practices with the nominal variable, sex, for the different age groups. For age groups 20-44, 45-64 and $\geq$65, we get $\hat{V} = .036$, $\hat{V} = .339$ and $\hat{V} = .347$, respectively which implies that the degree of association between grouped health practices and sex is stronger for the two older age groups. We must be somewhat cautious using Cramér's $V$ here, however, since the marginal distributions differ somewhat across the contingency tables for the three age groups. In addition, therefore, we examine these tables using the cross-product ratio which not only is not sensitive to the marginal distributions but also reveals underlying relationships in such a way that they are easily understood. From this we discover that there

is no difference in the number of health practices between males and females for the youngest age group. After examining several other measures of association, we conclude this regardless of which association measure we use. For the older age groups, however, there is a sex difference with females tending to have better health practices (as measured by the level of the grouped health practice score) than males. This tendency is strongest for the eldest age group where females are 4.6 times more likely than males to have the highest rather than the lowest health practice score. For the middle age group this odds ratio is 2.3. The difference between the sexes only appears when we compare those with the least number of health practices to those with a moderate or high number of practices. With odds ratios of approximately unity, no difference between males and females is apparent when the two highest levels of grouped health practices are compared.

The other three-way interaction in our model is between hospitalization, sex, and age ($HSA$). If we first look at the interaction between hospitalization and age for males and females, we note that the difference in the marginal distributions between the two tables is not so severe as to discard Cramér's $V$ as a comparative measure of magnitude of association. For males, $\hat{V} = .207$ while for females it is somewhat less at $\hat{V} = .103$. As before, however, we wish to look at additional measures.

Let us try to put a PRE interpretation on this relationship. There are several measures we could use in order to do so. We shall use Goodman and Kruskall's asymmetric measure $\tau$ where we treat hospitalization, $H$, as the dependent variable. As mentioned in our earlier section, Blalock (1972) and Reynolds (1977a) suggest $\tau$ over Goodman and Kruskall's PRE measure $\lambda$ when the marginal distribution of the dependent variable is highly skewed since $\tau$ is less sensitive to such skewness. In our case, although the marginal distributions are similar across the contingency tables for males and females, they are very highly skewed on the dependent variable,

hospitalization. For the cross-tabulation of hospitalization with age, the values of $\hat{\tau}_{II}$ are .043 and .011 for males and females, respectively. Given the proximity of these values to zero, we cannot say for either sex that knowledge of a person's age reduces the error in correctly predicting whether or not that person is hospitalized.

Although very close to zero, we recall that while independence of the variables implies that the measure is zero, the converse need not hold. A look at the cross-product ratio still uncovers interesting differences in behaviors between the sexes.

| sex | age | odds ratio (0 hospital days to $\geq$ 1 hospital days) |
|---|---|---|
| Male | 20-44 | 9.07 |
| | 45-64 | 3.96 |
| | $\geq$65 | 1.97 |
| Female | 20-44 | 3.60 |
| | 45-64 | 3.38 |
| | $\geq$65 | 1.84 |

For those in the two elder age groups, the odds of having no hospital days rather than at least one day in hospital are very similar for males and females. For these age groups, the odds decrease from 3.96 to 1.97 for males, and from 3.38 to 1.84 for females as we move from middle to old age. That is, for both sexes the odds of having zero rather than at least one hospital day, are approximately twice as high for those in the middle age group than for those in the eldest category. The difference in the sexes is manifested in the youngest age group with males 2.5 times more likely than females to have not been hospitalized at all as opposed to having spent at least one day in hospital.

It remains to look at the two-way interactions $AE$ and $PE$. Let us express $AE$ in

terms of a proportional reduction in error measure. Treating education as dependent upon age we note that since the marginal distribution of the dependent variable is highly skewed, we again use Goodman and Kruskall's $\tau$ measure. The proportional reduction in error resulting from moving from no knowledge of a person's age to knowledge of this independent variable, is $\hat{\tau}_E$=.120. Considering the cross-product ratios, we observe that young people are 3 times more likely than middle age people and almost 10 times more likely than old people, to have at least some post-secondary education rather than a maximum of a high school education; middle age people are over 3 times more likely than old people to have at least some post-secondary education rather than at most high school.

The final interaction in our model is that between grouped health practices and education, $PE$. The association between these two variables is not very strong. If we ignore those with the lowest grouped health practices score, then there is no difference in health practices for the two educational levels. On the other hand, if we include this low scoring group, we see that those with a low educational level are 1.7 times more likely than those in the higher educational level to have a low grouped health practice score versus a middle score; the low educational group is 1.9 times more likely than the higher educational group to have a low health practice score versus a high one.

The associations which exist in our model are discussed above. As mentioned in previous sections, there is no dependency between the health practice score and hospital utilization although we might expect that one's health habits would have some bearing on whether or not a person is hospitalized. As suggested by our analysis, it may be that this relationship does not exist, at least this simply. It may also be that the grouped health practice score as it is currently constructed for this analysis, is not a good health habit index measurement. This may be partly due to the individual

health practices being often too complex to dichotomize as either 'good' or 'bad'. This was already discussed in some detail. Another plausible explanation is that the health indicator variables which we have used here and which have been traditionally used in the literature are not only too complex to reduce to simple scores, but may not always be the best nor most appropriate variables for gauging health status. What may be required are additional variables and/or new constructions of current variables which are more appropriate for measuring one's level of health practices. The concept of a health practice score is a very complex one. More social medical study might be necessary to reassess and revise a better index.

# Chapter 4

# Conclusions

An effort was made primarily to ascertain if relationships exist between previously studied health indicator variables (eating breakfast, smoking, drinking, sleeping, weight, and exercise) and self-assessed health status, and between the health indicator variables and hospital utilization.

The analyses employed had the underlying assumption that the subjects were selected by means of a simple random sample. Instead, however, a single-stage cluster design was used to collect data on a sample from the adult population of Metropolitan St. John's, Newfoundland. For this reason design effects were calculated for two contingency tables known to be important to the analyses. Since in neither case were the design effects significant, the analytical techniques were used with confidence in their validity under the more complex sampling scheme.

Interactions between health habits and health status were explored using a variety of measures of association. Which measures were chosen reflected the particular contingency table being investigated, the information that was desired, whether or not the table was symmetric, whether the variables were nominal or ordinal, and so on. The strength of association varied with the health practice, with one's habit of eating breakfast being so weakly associated with health status that it was dropped as

a health indicator variable from the remainder of the analyses. The other indicator variables were more strongly related, although based on measures of association none were overwhelmingly related to health status.

Logistic regression was used to study in more detail the association between health status and sleeping, health status and drinking, and health status and drinking while controlling for education. Whereas sleeping and drinking were treated as grouped categorical variables in the contingency table analysis using measures of association, they were now treated as interval level, and health status was dichotomized as being either good or poor.

The logistic regression uncovered patterns to the association between these health practices and health status that were not apparent from the exploratory analysis. Examination of the relationships substantiated previous studies which showed that the frequency of reported good health status is optimum for those who sleep approximately 7 hours per night with the frequency declining for those with less than this and declining, but less dramatically, for those with more than 7 hours.

The association between drinking and health status was not as clearly defined as that between sleeping and health status. Even so, the logistic regression showed that, in general, people claimed to have good health less often as the amount of alcohol consumed increased; good health was most often reported by those who drank moderately or infrequently. The pattern of association changed somewhat once educational level was controlled for. Although it still held that moderate or infrequent drinking was best, this was most dramatically depicted for those with the highest level of education. In addition, it was clear from this analysis that, all else being equal, the higher the educational level, the more often health status was reported to be good.

Individual health indicator variables were studied and from these a weighted health practice index was constructed. We wished to see if health practices, as measured by a composite health habit index, are associated with hospital utilization once sex, age, and education level are controlled for. Loglinear analysis was used to build models to study the interactions between these variables from this five-way contingency table. Interesting interactions were uncovered with the most important, as reflected in our model, being between age and education, health practices and education, and between the two three-way interactions of health practices, sex, and age, and hospital utilization, sex, and age. The most notable interaction missing from our model was between the health practice score and hospitalizations; this was the interaction we had set out to examine in the loglinear analysis. One would expect a relationship between a health practice score and hospital utilization with people who have good health habits being hospitalized less frequently than those with poor health habits. Since this is not surfacing in this health study, it is suggested that the composition of the health index requires further study. In addition, it might be worthwhile to consider a more in-depth look at the hospitalization variable, both in terms of the frequency of and the reasons for the hospitalizations.

# References


Agresti, Alan (1984). *Analysis of Ordinal Categorical Data.* John Wiley & Sons, Inc., New York.

Altham, Patricia M.E. (1976). "Discrete Variable Analysis for Individuals Grouped into Families." *Biometrika*, 63(2):263–269.

Belloc, Nedra B. (1973). "Relationship of Health Practices and Mortality." *Preventative Medicine*, 2:67–81.

Belloc, Nedra B., and Lester Breslow (1972). "Relationship of Physical Health Status and Health Practices." *Preventative Medicine*, 1:409–421.

Belt, Jr., W. Thomas (1990). "Tobacco Smoking, Hypertension, Stroke, and Coronary Heart Disease: The Importance of Smoking Cessation." *Seminars in Respiratory Medicine*, 11(1) (January):36–49.

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice.* The Massachusetts Institute of Technology, Massachusetts.

Blalock, Jr., Hubert M. (1972). *Social Statistics.* McGraw Hill, New York, 2nd edition.

Breslow, Lester, and James E. Enstrom (1980). "Persistence of Health Habits and Their Relationship to Mortality." *Preventative Medicine*, 9:469–483.

Brier, Stephen S. (1980). "Analysis of Contingency Tables under Cluster Sampling." *Biometrika*, 67(3):591–596.

Brown, M.B., L. Engelman, M.A. Hill, and R.I. Jennrich (1988). *BMDP Statistical Software Manual, Volume 1.* The Regents of the University of California.

Brown, M.B., L. Engelman, M.A. Hill and R.I. Jennrich (1988). *BMDP Statistical Software Manual, Volume 2.* The Regents of the University of California.

Camilli, Gregory, and Kenneth D. Hopkins (1978). "Applicability of Chi-Square to 2×2 Contingency Tables with Small Expected Cell Frequencies." *Psychological Bulletin*, 85(1):163–167.

Cochran, William G. (1977). *Sampling Techniques.* John Wiley & Sons, Inc., New York, 3rd edition.

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, Jacob (1977). *Statistical Power Analysis for the Behavioral Sciences.* Academic Press, Inc., New York, revised edition.

Cohen, Joel E. (1976). "The Distribution of the Chi-Squared Statistic under Clustered Sampling from Contingency Tables." *Journal of the American Statistical Association*, 71(355) (September):665–670.

Conover, W.J. (1974). "Some Reasons for Not Using the Yates Continuity Correction on 2×2 Contingency Tables" (with comments). *Journal of the American Statistical Association*, 69(346) (June):374–382.

Cook, D.G., S.J. Pocock, A.G. Shaper, and S.J. Kussick (1986). "Giving up Smoking and the Risk of Heart Attacks." *The Lancet*, 2 (December):1376–1379.

Costner, Herbert L. (1965). "Criteria for Measures of Association." *American Sociological Review*, 30(3) (June):341–353.

Cramér, Harald (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

Davis, James A. (1971). *Elementary Survey Analysis.* Prentice-Hall Methods of Social Science Series. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Everitt, B.S. (1977). *The Analysis of Contingency Tables.* Chapman and Hall, Inc., New York.

Fellegi, Ivan P. (1978). "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multi-Stage Samples." *Ottawa: Survey Methodology, Statistics Canada*, 4(1):29–56.

Fellegi, Ivan P. (1980). "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples." *Journal of the American Statistical Association*, 75(370) (June):261–268.

Fienberg, Steven E. (1977). *The Analysis of Cross-Classified Categorical Data.* The Massachusetts Institute of Technology, Massachusetts.

Fingleton, B. (1984). *Models of Category Counts.* Cambridge University Press, Cambridge.

Fisher, Sir Ronald A. (1970). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, 14th edition.

Freeman, Jr., Daniel H. (1987). *Applied Categorical Data Analysis.* Statistics: Textbooks and Monographs. Volume 79. Marcel Dekker, Inc., New York.

Garson, G. David (1976). *Political Science Methods.* Holbrook Press, Inc., Boston.

Gart, John J. (1971). "The Comparison of Proportions: A Review of Significance Tests, Confidence Intervals and Adjustments for Stratification." *Review of the International Statistical Institute*, 39(2):148–169.

Gart, John J., and James R. Zweifel (1967). "On the Bias of Various Estimators of the Logit and its Variance with Application to Quantal Bioassay." *Biometrika*, 54:181–187.

Goodman, Leo A. (1968). "The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interactions in Contingency Tables with or without Missing Entries." *Journal of the American Statistical Association*, 63(324) (December):1091–1131.

Goodman, Leo A. (1969). "How to Ransack Social Mobility Tables and Other Kinds of Cross-Classification Tables." *American Journal of Sociology*, 75(1) (July):1–40.

Goodman, Leo A. (1978). *Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent-Structure Analysis.* Abt Associates, Inc., Cambridge, Massachusetts.

Goodman, Leo A., and William H. Kruskal (1954). "Measures of Association for Cross Classifications." *Journal of the American Statistical Association*, 49(268): 732–764.

Goodman, Leo A., and William H. Kruskal (1963). "Measures of Association for Cross Classifications III: Approximate Sampling Theory." *Journal of the American Statistical Association*, 58 (June):310–364.

Goodman, Leo A., and William H. Kruskal (1972). "Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances." *Journal of the American Statistical Association*, 67(338) (June):415–421.

Goodman, Leo A., and William H. Kruskal (1979). *Measures of Association for Cross Classifications.* Springer-Verlag, New York. (Contains four articles including the two just listed).

Griffith, David E., and Joe G.N. Garcia (1989). "Tobacco Cigarettes, Smoking, Smoking Cessation, and Chronic Obstructive Pulmonary Disease." *Seminars in Respiratory Medicine*, 10(4) (October):356–371.

Grizzle, James E. (1967). "Continuity Correction in the $\chi^2$-Test for 2×2 Tables." *The American Statistician*, 21(4) (October):28–32. The Teacher's Corner.

Hays, William L. (1981). *Statistics.* CBS College Publishing, New York, 3rd edition.

Holt, D., A.J. Scott, and P.D. Ewings (1980). "Chi-Squared Tests with Survey Data." *Journal of the Royal Statistical Society*, A, Part 3(143):303–320.

Hosmer, Jr., David W., and Stanley Lemeshow (1989). *Applied Logistic Regression.* John Wiley & Sons, Inc., New York.

Irwin, J. O. (1949). "A Note on the Subdivision of $\chi^2$ into Components." *Biometrika*, 36:130–134.

Iverson, Gudmund R. (1979). "Decomposing Chi-Square: A Forgotten Technique." *Sociological Methods & Research*, 8(2) (November):143–157.

Kendall, Sir Maurice, and Alan Stuart (1979). *The Advanced Theory of Statistics*, volume 2. Charles Griffin & Company Limited, London & High Wycombe, 4th edition.

Kimball, A. W. (1954). "Short-Cut Formulas for the Exact Partition of $\chi^2$ in Contingency Tables." *Biometrics*, 10 (December):452–458.

Kraemer, Helena Chmura, and Sue Thiemann (1987). *How Many Subjects? Statistical Power Analysis in Research.* Sage Publications, Inc., Newbury Park, California.

Lancaster, H. O. (1949). "The Derivation and Partition of $\chi^2$ in Certain Discrete Distributions." *Biometrika*, 36:117–129.

Lavrakas, Paul J. (1987). *Telephone Survey Methods: Sampling, Selection and Supervision.* Applied Social Research Methods Series. Volume 7. Sage Publications, Inc., Newbury Park, California.

Liebetrau, Albert M. (1983). *Measures of Association.* Sage University Papers series on Quantitative Applications in the Social Sciences, 07-032. Sage Publications, Inc., Beverly Hills, California.

Maxwell, A.E. (1961). *Analysing Qualitative Data.* Methuen & Co. Ltd., London.

Maxwell, A.E. (1978). *Basic Statistics for Medical and Social Science Students.* Chapman and Hall, London.

Miller, G.H., Charles E. Chittenden, and Robert J. Myers (1990). "Life Expectancy at Age 30: Nonsmoking Versus Smoking Men." *Contingencies*, pages 30–45, (May/June).

Minitab Inc. (1989). *Minitab Reference Manual, Release 7.*

Mueller, John H., Karl F. Schuessler, and Herbert L. Costner (1977). *Statistical Reasoning in Sociology.* Houghton Mifflin Company, Boston, 3rd edition.

Norušis, Marija J. (1985). *SPSS$^x$ Advanced Statistics Guide.* SPSS Inc., Chicago.

Pielou, E. C. (1969). *An Introduction to Mathematical Ecology.* John Wiley & Sons, Inc., New York.

Plackett, R. L. (1981). *The Analysis of Categorical Data.* Griffin's Statistical Monograph Series. Monograph No. 35. Macmillan Publishing Co., Inc., New York, 2nd edition.

Potthoff, Richard F. (1987). "Some Generalizations of the Mitofsky-Waksberg Technique for Random Digit Dialing." *Journal of the American Statistical Association*, 82(398) (June):409–418.

Rao, J.N.K., and A.J. Scott (1981). "The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables." *Journal of the American Statistical Association*, 76(374) (June):221–230.

Rao, J.N.K., and A.J. Scott (1984). "On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data." *The Annals of Statistics*, 12(1):46–60.

Reynolds, H.T. (1977a). *The Analysis of Cross-Classifications.* The Free Press, A Division of Macmillan Publishing Co., Inc., New York.

Reynolds, H.T. (1977b). *Analysis of Nominal Data.* Sage University Papers series on Quantitative Applications in the Social Sciences, 07-007. Sage Publications, Inc., Beverly Hills, California.

Segovia, Jorge, Roy Bartlett, Barbara Veitch, and Alison Edwards (1986a). "The St. John's Study of Health Practices and Medical Care Utilization; Medical Care Contributed Papers V." American Public Health Association, (June). 114th Annual Meeting, Las Vegas.

Segovia, Jorge, Roy Bartlett, Barbara Veitch, and Alison Edwards (1986b). "Health Practices, Health Status and Medical Care Utilization: A Preliminary Report." Canadian Public Health Association, (September). 77th Annual Conference, Vancouver.

Segovia, Jorge, Roy Bartlett, Barbara Veitch, and Alison Edwards (1987). *Lifestyles, Health Practices and Utilization of Health Services.* Faculty of Medicine, Memorial University of Newfoundland.

SPSS Inc. (1986). Chicago. *SPSS Graphics.*

SPSS Inc. (1988). Chicago. *SPSS-X User's Guide*, 3rd edition.

Somers, Robert H. (1962). "A New Asymmetric Measure of Association for Ordinal Variables." *American Sociological Review*, 27(6):799–811.

Stuart, Alan (1953). "The Estimation and Comparison of Strengths of Association in Contingency Tables." *Biometrika*, 40:105–110.

Sudman, Seymour (1976). *Applied Sampling.* Quantitative Studies in Social Relations. Academic Press, Inc., New York.

Thomas, D. Roland, and J.N.K. Rao (1987). "Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics under Cluster Sampling." *Journal of the American Statistical Association*, 82(398) (June):630–636.

Upton, Graham J.G. (1978). *The Analysis of Cross-tabulated Data.* John Wiley & Sons Ltd., Chichester.

U.S. Department of Health and Human Services (1990). *The Health Benefits of Smoking Cessation. A Report of the Surgeon General.* U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office of Smoking and Health. DHHS Publication No. (CDC) 90-8416, (September).

Waksberg, Joseph (1978). "Sampling Methods for Random Digit Dialing." *Journal of the American Statistical Association*, 73(361) (March):40–46.

Warner, Mark A., et al. (1989). "Role of Preoperative Cessation of Smoking and Other Factors in Postoperative Pulmonary Complications: A Blinded Prospective Study of Coronary Artery Bypass Patients." *Mayo Clinic Proceedings*, 64 (June):609–616.

Wilson, Thomas P. (1974). "Measures of Association for Bivariate Ordinal Hypotheses." *Measurement in the Social Sciences*, edited by Hubert M. Blalock, Jr., Aldine Publishing Company, Chicago, pp. 327–342.

Yates, F. (1934). "Contingency Tables Involving Small Numbers and the $\chi^2$ Test." *Journal of the Royal Statistical Society, Supplement*, 1(2):217–235.

Yule, G.U. (1912). "On the Methods of Measuring Association between Two Attributes." *Journal of the Royal Statistical Society*, 75 (May):579–642.

# Appendix

## Questionnaire

| | |
|---|---|
| INTERVIEWER | ☐ |
| HOUSEHOLD | ☐☐☐☐ |
| SUBJECT | ☐☐☐ |
| DATE OF INTERVIEW | ☐☐☐☐☐☐ |
| DATE RECEIVED | ☐☐☐☐☐☐ |
| BATCH NO. | ☐☐ |

Memorial University of Newfoundland
Faculty of Medicine
DIVISION OF COMMUNITY MEDICINE AND
BEHAVIOURAL SCIENCES

# LIFESTYLE, HEALTH PRACTICES AND MEDICAL CARE UTILIZATION SURVEY QUESTIONNAIRE

** TO BE REMOVED BY FIELD OFFICE BEFORE DATA PROCESSING

SUBJECT'S NAME_____TELEPHONE NO._____

ADDRESS_____

CONSENT STATEMENT READ ☐

Int. ☐ 1

I.D. Household ☐☐☐☐ 2

Subject ☐☐☐ 3

(M.C.P. NUMBER) FROM Q. 65 ☐☐☐☐☐☐☐☐☐☐☐☐ 4

**WE ARE GOING TO ASK ABOUT YOUR EATING AND SLEEPING HABITS NOW.**

1. How often do you eat breakfast?
   PROBE: *(EATING BREAKFAST MEANS MORE THAN A CUP OF COFFEE ONLY)*

   Every day, or almost every day 1 ☐
   Sometimes (3-4 times a week) 2 ☐
   Rarely, or never 3 ☐

   ☐ 5

2. Do you make any conscious effort to limit the amount of red meat in your diet for *health reasons?*

   Yes 1 ☐
   No 2 ☐

   ☐ 6

3. Do you make any conscious effort to limit the amount of animal fat in your diet?
   PROBE: (REAL BUTTER, WHOLE MILK, EGGS)

   Yes 1 ☐
   No 2 ☐

   ☐ 7

4. How many hours do you sleep per night?
   (PROBE: SIX, SEVEN?) *CODE NUMBER OF HOURS DIRECT*

   ☐☐ ☐☐ 8

5. How tall are you?
   *(WRITE THE ANSWER IN THE UNITS GIVEN BY THE RESPONDENT)*

   ☐ feet ☐☐ • ☐ ins cm ☐☐☐ ☐☐☐ 9

6. How much do you weigh?
   *RECORD THE ANSWER IN THE UNITS GIVEN BY RESPONDENT*

   ☐☐☐ lbs kg ☐☐☐ ☐☐☐ 10

7. Do you consider yourself to be . . . .

   . . . overweight 1 ☐
   . . . underweight 2 ☐
   . . . about average 3 ☐
   . . . DK 9 ☐

   ☐ 11

8. You are . . . Male 1 ☐ Female 2 ☐

   ☐ 12

**THE NEXT QUESTIONS ARE ABOUT SMOKING AND DRINKING**

9. Did you ever smoke regularly?

   PROBE: REGULAR SMOKING MEANS ONE CIGARETTE, PIPE, CIGAR A DAY FOR ONE YEAR

   YES 1 ☐ NO 2 ☐ *GO TO Q. 18* → 13

10. Are you smoking now?

    YES 1 ☐ NO 2 ☐ *GO TO Q. 12* → 14

11. Do you smoke . . . . . CURRENT

| | Yes | No | |
|---|---|---|---|
| . . . . . cigarettes | 1 ☐ | 2 ☐ | 15 |
| . . . . . pipe | 1 ☐ | 2 ☐ | 16 |
| . . . . . cigar | 1 ☐ | 2 ☐ | 17 |

12. Did you ever regularly smoke . . . *ASK WHATEVER NOT MENTIONED ABOVE*

| | EX | NEVER | |
|---|---|---|---|
| . . . . . cigarettes | 1 ☐ | 2 ☐ | 18 |
| . . . . . pipe | 1 ☐ | 2 ☐ | 19 |
| . . . . . cigar | 1 ☐ | 2 ☐ | 20 |

*IF YES TO ANY OF THE ABOVE*

13. How old were you when you stopped smoking? *THE ABOVE*

| | | | |
|---|---|---|---|
| . . . . . cigarettes | ☐☐ | *CODE AGE DIRECT* NA 99 | 21 |
| . . . . . pipe | ☐☐ | | 22 |
| . . . . . cigar | ☐☐ | | 23 |

*ASK ALL SMOKERS PAST AND PRESENT*

14. How old were you when you started smoking? *ASK WHATEVER MENTIONED*

| | | | |
|---|---|---|---|
| . . . . . cigarettes | ☐☐ | *CODE AGE DIRECT* NA 99 | 24 |
| . . . . . pipe | ☐☐ | | 25 |
| . . . . . cigar | ☐☐ | | 26 |

15. During the period when you smoked most, how many cig/pipes.cigars/did you smoke a day?

| | | | |
|---|---|---|---|
| . . . . . cigarettes | ☐☐ | *CODE AGE DIRECT* NA 99 | 27 |
| . . . . . pipe | ☐☐ | | 28 |
| . . . . . cigars | ☐☐ | | 29 |

16. Do/Did you inhale the smoke?

| | Yes | No | | |
|---|---|---|---|---|
| . . . . . cigarettes | 1 ☐ | 2 ☐ | NA 99 | 30 |
| . . . . . pipe | 1 ☐ | 2 ☐ | | 31 |
| . . . . . cigar | 1 ☐ | 2 ☐ | | 32 |

*ASK CURRENT SMOKERS*

17. During the past two years, did you make a serious attempt to stop smoking?

    Yes 1 ☐ No 2 ☐ NA 9 → 33

18. Do you drink any alcoholic beverages, that is beer, wine or liquor?

Yes 1 ☐ No 2 ☐ 34

19. Did you ever drink alcoholic beverages once a month or more?

Yes 1 ☐ No 2 ☐ 35

*GO TO Q. 25*

20. On the average how often do you drink alcoholic beverages such as beer, wine or liquor?

On the average, how often did you drink alcoholic beverages such as beer, wine or liquor?

| | | |
|---|---|---|
| Every day | 1 | ☐ |
| 5-6 days a week | 2 | ☐ |
| 3-4 days a week | 3 | ☐ |
| 1-2 days a week | 4 | ☐ 36 |
| 2-3 times a month | 5 | ☐ |
| Once a month | 6 | ☐ |
| Less than once a month | 7 | ☐ |

21. On the days you drink, about how many drinks do you have per day?

*CODE DIRECT* ☐☐

22. Have you recently (in the past 6 months) changed your drinking habits because of a health problem?

Yes 1 ☐ No 2 ☐ N.A. 9

On the days you drank about how many drinks did you have per day?

*CODE DIRECT* ☐☐ 37

When did you stop drinking?

*CODE YEAR DIRECT* 38

23. ☐☐ 39

Did you stop for health reasons?

24. Yes 1 ☐ No 2 ☐ 40

**THE NEXT SECTION IS ABOUT YOUR PHYSICAL ACTIVITIES**

25. Are you now suffering from any disability

(PROBE: A CONDITION THAT STOPS YOU FROM DOING YOUR ROUTINE ACTIVITIES)

Yes 1 ☐
No 2 ☐ *GO TO Q. 29* → ☐ 41

26. Is it a temporary condition?

(PROBE: A CONDITION THAT WILL DISAPPEAR IN A FEW WEEKS)

Yes 1 ☐
No 2 ☐ → ☐ 42
DK 9 ☐

27. Was it caused by an accident or injury?

Yes 1 ☐
No 2 ☐ *GO TO Q 29* → ☐ 43

28. Did this accident or injury happen . . . .

. . . . at home 1 ☐
. . . . outdoors 2 ☐
. . . . traffic 3 ☐ → ☐ 44
. . . . at work 4 ☐

29. How many times in a 2 week period do you usually do any of the following exercises or recreational activities?

How much time did you spend on each occasion?

| | *READ* | No. of Times | Mins 1-10 | Mins 15+ | NA | | | |
|---|---|---|---|---|---|---|---|---|
| 1. | Walking (including to and from school or work) | ☐☐ | 1 | 2 | 9 | 45 | ☐☐ | 46 ☐ |
| 2. | Jogging or running | ☐☐ | 1 | 2 | 9 | 47 | ☐☐ | 48 ☐ |
| 3. | Calisthenics (doing physical exercises) | ☐☐ | 1 | 2 | 9 | 49 | ☐☐ | 50 ☐ |
| 4. | Bicycling (including to and from work) | ☐☐ | 1 | 2 | 9 | 51 | ☐☐ | 52 ☐ |
| 5. | Bowling | ☐☐ | 1 | 2 | 9 | 53 | ☐☐ | 54 ☐ |
| 6. | Vigorous dancing | ☐☐ | 1 | 2 | 9 | 55 | ☐☐ | 56 ☐ |
| 7. | Skating | ☐☐ | 1 | 2 | 9 | 57 | ☐☐ | 58 ☐ |
| 8. | Team sports (such as baseball, softball etc) | ☐☐ | 1 | 2 | 9 | 59 | ☐☐ | 60 ☐ |
| 9. | Swimming | ☐☐ | 1 | 2 | 9 | 61 | ☐☐ | 62 ☐ |
| 10. | Gardening | ☐☐ | 1 | 2 | 9 | 63 | ☐☐ | 64 ☐ |
| 11. | Racquet sports | ☐☐ | 1 | 2 | 9 | 65 | ☐☐ | 66 ☐ |
| 12. | Golf | ☐☐ | 1 | 2 | 9 | 67 | ☐☐ | 68 ☐ |
| 13. | Other (Specify) | ☐☐ | 1 | 2 | 9 | 69 | ☐☐ | 70 ☐ |

30. Are you more, less, or equally active in winter?

More 1 ☐
Less 2 ☐
Equally 3 ☐ 71

**AND NOW SOME QUESTIONS IN RELATION TO MEDICAL CARE**

31. In the last year, that is from of 1984, did you have a consultation with a doctor?

Yes 1 ☐
No 2 ☐ *Go to Q 33* 72

32. How many visits did you have in the last year?

*CODE DIRECT*
*NA or DK 99* ☐☐ 73

33. Do you have a family doctor?

(PROBE: A DOCTOR WHOM YOU ALWAYS CONSULT)

Yes 1 ☐
No 2 ☐ 74

34. Within the last year (from 1984) have you stayed at home because of an illness, or not feeling well

Yes 1 ☐
No 2 ☐ *GO TO Q. 37* 75

35. Did you stay in bed?

Yes 1 ☐
No 2 ☐ 76

36. How many days did you stay in bed?

*CODE DIRECT* ☐☐☐ 77

37. In the last year (same period) have you been a patient in a hospital overnight?

(PROBE: DID YOU SPEND AT LEAST A NIGHT IN A HOSPITAL)

Yes 1 ☐
No 2 ☐ *GO TO Q. 40* 78

38. How many days did you spend at the hospital?

*CODE DIRECT* ☐☐☐ 79

*FOR FEMALES ONLY*

39. Was the hospitalization due to pregnancy or delivery?

Yes 1 ☐
No 2 ☐ 80

40. We would like to know how satisfied or dissatisfied you are, in general with medical care in your own experience. On a five-point scale in which 5 means that you are very satisfied, and 1 means that you are very dissatisfied, what will be your score?

(PROBE: THINK OF A LADDER WITH FIVE RUNGS, WHERE THE HIGHEST OF THE FIVE IS THE BEST, WHERE ARE YOU ON THIS LADDER?)

SATISFIED

TICK

| | |
|---|---|
| 5 | |
| 4 | |
| 3 | |
| 2 | |
| 1 | |

DISSATISFIED

DK 9 ☐ ☐ 81

**NOW LETS GO BACK TO YOUR OWN HEALTH AND WELL BEING**

41. Would you say that your health is . . . .

| | | |
|---|---|---|
| . . . . Excellent | 1 | ☐ |
| . . . . Good | 2 | ☐ |
| . . . . Fair | 3 | ☐ |
| . . . . Poor | 4 | ☐ |

☐ 82

42. Over the past year, has your health caused you. . . .

| | | |
|---|---|---|
| . . . . no worry at all | 1 | ☐ |
| . . . . Hardly any worry | 2 | ☐ |
| . . . . Some worry | 3 | ☐ |
| . . . . A great deal of worry | 4 | ☐ |

☐ 83

43. Do you have any of the following chronic conditions?

(CHRONIC MEANS THE CONDITION HAS BEEN PRESENT FOR THREE MONTHS OR MORE)

*READ LIST:*

*CIRCLE CODES THAT CORRESPOND*

| | | | |
|---|---|---|---|
| Anemia | 01 | High blood pressure | 13 |
| Allergy (OF ANY KIND) | 02 | Kidney disease (stones etc.) | 14 |
| Arthritis, rheumatism | 03 | Mental illness | 15 |
| Asthma | 04 | Missing arm(s) or leg(s) | 16 |
| Cancer | 05 | Missing finger(s) toes | 17 |
| Cerebral Palsy | 06 | Paralysis of any kind | 18 |
| Diabetes | 07 | MALES: Prostrate disease | 19 |
| FEMALES: Dysmenorrhea | | Recurring backaches | 20 |
| (menstrual problems) | 08 | Recurring headaches | 21 |
| Emphysema | 09 | Stomach ulcer | 22 |
| Epilepsy | 10 | Thyroid trouble or goitre | 23 |
| Heart disease | 11 | Tuberculosis | 24 |
| Hemorrhoids (piles) | 12 | OTHER | |
| | | Specify_____ | 25 |
| | | None | 88 |

☐☐ 84
☐☐ 85
☐☐ 86
☐☐ 87
☐☐ 88
☐☐ 89
☐☐ 90
☐☐ 91

44. Compared with other people your age, would you say you have . . . . .

| | | |
|---|---|---|
| . . . . much more energy | 1 | |
| . . . . somewhat more (energy) | 2 | |
| . . . . average amount of energy | 3 | |
| . . . . somewhat less (energy) | 4 | |
| . . . . much less energy | 5 | |

92

45. In general, how satisfied are you with your overall physical condition. . . .

| | | |
|---|---|---|
| . . . . are you very satisfied | 1 | |
| . . . . satisfied | 2 | |
| . . . . not too satisfied | 3 | |
| . . . . not at all satisfied | 4 | |

93

46. During the past few weeks, how often have you felt. . . .
*CIRCLE*

| would you say . . . . . | Often | Sometimes | Never | |
|---|---|---|---|---|
| . . . . on top of the world | 1 | 2 | 3 | 94 |
| . . . . lonely | 1 | 2 | 3 | 95 |
| . . . . that things were going your way | 1 | 2 | 3 | 96 |
| . . . . restless | 1 | 2 | 3 | 97 |
| . . . . depressed, or unhappy | 1 | 2 | 3 | 98 |

47. All in all, how happy are you these days? Would you say. . .

| | | |
|---|---|---|
| . . . . very happy | 1 | |
| . . . . pretty happy | 2 | |
| . . . . not too happy | 3 | |
| . . . . unhappy | 4 | |

99

48. How many close relative do you have? These are people that you feel at ease with, can talk to about private matters, and can call on for help. *(DO NOT INCLUDE SPOUSE)*

*CODE DIRECT*

100

49. How many close friends do you have? These are people that you feel at ease with, can talk to about private matters and can call on for help.

*CODE DIRECT*

101

**NOW SOME QUESTIONS ABOUT PREVENTIVE HEALTH:**

50. How often do you brush your teeth?
PROBE (DO YOU HAVE YOUR OWN TEETH?)

PROBE: More than twice a day 1 ☐
Twice a day 2 ☐
Once a day 3 ☐
Not every day 4 ☐
No teeth Code NA 9 ☐

*GO TO Q. 54* → ☐ 102

51. Do you use dental floss? (WATER PICK COUNTS AS FLOSS)

Yes 1 ☐
No 2 ☐ *GO TO Q. 53* → ☐ 103

52. How often?

Every day 1 ☐
Every week 2 ☐ → ☐ 104

53. When was the last time that you went to a dentist?

USE LIST AS PROBE:
Within the last year 1 ☐
one to two years 2 ☐
more than two years 3 ☐
Never 4 ☐
DK 9 ☐ → ☐ 105

54. When was the last time that your went to a doctor for a preventive examination when you were not sick?

PROBE: FEMALES: PAP SMEAR, BREAST EXAMINATION
MALES: BLOOD PRESSURE CHECK

*CODE YEAR DIRECT*

☐☐ → ☐☐ 106

Never 00
DK 99

55. Do you use your seatbelt while travelling by car?

Yes 1 ☐
No 2 ☐ → ☐ 107

**TO COMPLETE THE QUESTIONNAIRE WE NEED A FEW MORE DETAILS:**

56. Where were you born? Was it Newfoundland?

*IF CANADA, ASK PROVINCE. IF NOT IN CANADA ASK COUNTRY. CIRCLE*

| | | | | | |
|---|---|---|---|---|---|
| NFLD. | 01 | MAN. | 07 | U.K. | 13 |
| N.S. | 02 | SASK. | 08 | U.S.A. | 14 |
| N.B. | 03 | ALBTA | 09 | OTHER | |
| P.E.I. | 04 | B.C. | 10 | AMERICAS | 15 |
| QUE. | 05 | YUKON | 11 | EUROPE | 16 |
| ONTARIO | 06 | N.W.T. | 12 | ASIA | 17 |
| | | | | OTHER | 18 |

☐☐ 108

57. What is your marital status?
PROBE: ARE YOU MARRIED?

| | |
|---|---|
| Single | 1 ☐ |
| Married | 2 ☐ |
| Divorced/Separated | 3 ☐ |
| Widowed | 4 ☐ |

☐ 109

58. What was the last grade you complete in school?

*CODE DIRECT* ☐☐

☐☐ 110

59. ASK ONLY IF ANSWER INDICATES THAT RESPONDENT COMPLETED HIGH SCHOOL

Do you have any education beyond High School?

| | | |
|---|---|---|
| Yes | 1 ☐ | |
| No | 2 ☐ | *GO TO Q. 62* |

☐ 111

60. What kind of education was it?

| | | |
|---|---|---|
| Tradeschool, diploma courses etc. | 1 ☐ | *GO TO Q. 62* |
| University | 2 ☐ | |

☐ 112

61. Do you have a university degree?

| | | |
|---|---|---|
| Yes | 1 ☐ | |
| No | 2 ☐ | |

☐ 113

62. Are you now.... working ..... 1
.... retired ..... 2
.... unemployed ..... 3
.... laid off temp/on strike ..... 4
.... unable to work (disability) ..... 5
.... keeping house ..... 6
.... studying ..... 7

*CIRCLE*

*GO TO Q. 64*

☐ 114

63. What is/was your job?
PROBE: WHAT DO YOU DO AT WORK?

_____ 115

64. What is your date of birth? Y Y M M D D 116

65. What is your M.C.P. No.?

ASK HUSBAND AND WIFE ONLY

66. What is the approximate total income for your household?
(PROBE: INCLUDING ALL WAGES, SALARIES, PENSIONS, AND ALLOWANCES)

. . . . is it less than $15,000 1
. . . . between $15,000 and $30,000 2
. . . . more than $30,000 3
N.A. 9 119

**THAT COMPLETES THE INTERVIEW. THANK YOU VERY MUCH FOR DONATING YOUR TIME TO THE STUDY. IT IS VERY MUCH APPRECIATED.**

TO BE CODED FOLLOWING COMPLETION OF INTERVIEW, FROM THE HOUSEHOLD SHEET

CODE EITHER TO WIFE (OR SINGLE FEMALE) OR HUSBAND (OR SINGLE MALE)

CODE 9 for all the rest — DO NOT LEAVE BLANKS

| | Deceased 1 | Independent 2 | Family 3 | Nursing H. 4 | H.H. 5 | |
|---|---|---|---|---|---|---|
| 67. WIFE'S Mother | | | | | | 120 |
| Father | | | | | | 121 |
| 68. HUSBAND'S Mother | | | | | | 122 |
| Father | | | | | | 123 |

CODE FROM HOUSEHOLD SHEET, AFTER COMPLETION OF ALL INTERVIEWS

69. Total number of subjects in H.H. 124
Total number of children 19 or less 125

TO BE CODED ONLY BY FIELD OFFICE:

70. Total number of refusals 126
Total number of non-respondents 127